Univerzita Karlova v Praze

Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Učitelství češtiny jako cizího jazyka

Diplomová práce

Bc. Martina Vokáčová

Vliv gramatických profilů českých substantiv na jejich osvojování nerodilými mluvčími

The influence of grammatical profiles of Czech nouns on their acquisition by non-native speakers

Praha 2016

Vedoucí práce: Mgr. Eva Lehečková, Ph.D.

Děkuji vedoucí práce, Mgr. Evě Lehečkové, Ph.D., za těžko shrnutelnou spoustu věcí. Například ale za léto s gramatickými profily. Za to, že tahle práce vznikla. Za zaujetí. Děkuji také svým přátelům – Janě, Martinovi a Anče za pomoc s anotací dat, Ivanovi za souputnictví, Olince za kontrolování stavu textu i jeho autorky a Karolíně za Tofu.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 16. srpna 2016 .....

Klíčová slova:
gramatické profily, frekvence, substantiva, osvojování druhého jazyka, CZESL

Keywords:

grammatical profiles, frequency, nouns, second language acquisition, CZESL

Abstrakt:

Předkládaná práce zkoumá vliv frekvenčních charakteristik českých substantiv na jejich osvojování nerodilými mluvčími. První, teoretická část shrnuje diskuzi o významu frekvence pro reprezentaci gramatických kategorií v mysli, která se odehrává v rámci zahraniční kognitivní lingvistiky a psycholingvistiky. Druhá část představuje metodiku výzkumu, způsob získávání a anotace vzorku 20 lemmat z korpusu nerodilých mluvčích češtiny CzeSL-SGT. Ve třetí části je provedena analýza vybraných substantiv s ohledem na jejich gramatické profily (tvořené dvěma až třemi v korpusu SYN2015 nejfrekventovanějšími pádovými tvary). Ukazuje, že produkce nerodilých mluvčích má tendenci se od gramatických profilů odvíjet, což se projevuje jednak ve vysoké korespondenci frekvenčních charakteristik substantiv, jednak v nízkém zastoupení morfologických chyb v jejich nejčetnějších tvarech. Případy, kdy se chybovost od tohoto modelu odchyluje, vysvětluje pomocí typové frekvence – souběžně působícího efektu produktivity určitých deklinačních vzorů – a vyšší relevancí nominativu, jakožto základního tvaru, pro nerodilé mluvčí.

Abstract:

The present thesis examines the influence of Czech nouns frequency features on their acquisition by non-native speakers. The first theoretical part summarizes the ongoing discussion regarding the importance of frequency for the entrenchment of grammatical categories in one's mind as outlined by cognitive and psycho-linguists abroad. The second introduces the research methodology, collection method and annotation of 20 lemmas obtained from non-native Czech speaker corpus CzeSL-SGT. Subsequently, an analysis of selected nouns is carried out with regard to their grammatical profiles (comprising two or three most frequent case forms as found in SYN2015). Based on its results, the production of non-native speakers shows a tendency to follow grammatical profiles as demonstrated by high correspondence of frequency features of nouns on the one hand and by low error rate in morphology of most frequent forms on the other. Additionally, cases where lower rate does not correspond to the above mentioned model can be explained using type frequency, i.e. the productivity of particular declension models having simultaneously more far-reaching effect, and higher relevance of nominative as the default form for non-native speakers.

OBSAH

Motto

Minulý týden jsem napsala diplomce

(z žákovského korpusu Schöne 2015)

Mnoha slova a deklinace jsou podobné. Ale, když začneš žít češtinou, tak chápáš, že čeština je celý svět

(z korpusu CzeSL)

1 Úvod

Lída Holá, autorka a spoluautorka v současnosti pravděpodobně nejužívanějších učebnic češtiny pro cizince (řady Čeština expres a Czech step by step) publikovala na metodickém portálu k těmto učebnicím úvahu na téma, zda se čeština dá číst odzadu.[1] Pod tímto názvem se skrývá apel na učitele češtiny jako cizího jazyka, aby ve výuce přestali upozaďovat nominativ. Pro situaci, kdy by student češtiny jako cizího jazyka četl česká slova odzadu, předpokládá, že:

„Česká slova bude luštit "odzadu", tedy pomocí koncovek, určitě hezky dlouho: akuzativ se mu bude plést s genitivem, dativ s lokálem, a dokonce i lehký instrumentál a tu i onde odposlouchaný vokativ mu připraví nejednu horkou chvilku... Ale počkejme, přemýšlí takový svědomitý student, není těch pádů nějak málo? Ovšem, málem jsme zapomněli na nominativ. A právě zde skrývá kámen úrazu. My, čeští učitelé pro cizince, totiž pokládáme nominativní tvar podstatných jmen za tak samozřejmý a jednoduchý, že ho studentům často předkládáme jako jakýsi "pád bez koncovky". Mají ho ve slovníku, musejí se ho naučit a hotovo. "První pád" se tak stává pádem posledním, jakousi všední popelkou mezi pády.“

Nadsazený myšlenkový experiment Lídy Holé obsahuje řadu predikcí o tom, v jakých tvarech student češtiny pravděpodobně bude chybovat, na jakém základě k chybě dojde a s čím se často či nejméně často ve výuce setkává.

Cílem předkládané práce je v podstatě odpovědět na otázku, jak se to s oním modelovým přemýšlivým studentem má „ve skutečnosti“ a jaký pád lze označit za popelku. Jinými slovy se v této práci budu zabývat otázkou vlivu frekvenčních charakteristik / gramatických profilů českých substantiv na jejich osvojování nerodilými mluvčími. Gramatický profil substantiva tvoří sestupné pořadí jeho nejfrekventovanějších tvarů. Předpokládá se, že substantiva se stejným gramatickým profilem se shlukují do tříd podobných sémantických a funkčních rysů. Na základě korpusové sondy budu sledovat, zda jsou gramatické profily substantiv pro nerodilé mluvčí kognitivně relevantní entitou a odráží se v produkci nerodilých mluvčích, tj.

[1] Dostupné z: http://www.czechstepbystep.cz/clanky/da_se_cestina_cist_od_zadu.html, cit. 15. 8. 2016.

zda jsou nejčetnější tvary určitého substantivního lemmatu produkovány s nižším procentem chybovosti.

Materiálově práce čerpá z korpusů nerodilých mluvčích CzeSL a Merlin a z výzkumu Lehečkové, Lázničky a Jandy (2016) přebírá profily substantiv získaných v SYN2015. Analýza bude obsahovat dvě části. Jednat analyzuji chybovost podle vzoru pilotního výzkumu Karin Schöne (2015) a ověřím tak platnost jejích závěrů pro větší vzorek morfologicky chybných tvarů, jednak se budu věnovat analýze gramatických profilů a jejich vlivu na produkované chyby.

Práce je strukturována následovně. V kapitole 2 představím vybrané teoretické pohledy na roli frekvence v jazyce a jeho osvojování; v kapitole 3 popíšu metodiku výzkumu, sestavování vzorku a zásady jeho anotace; v kapitole 4 tento vzorek analyzuji a výsledky analýzy shrnuji v závěrečné části 5, kde také nastíním možnosti dalšího vytěžení tohoto tématu.

2 Frekvence a SLA

2.1 Proč počítat s frekvencí?

Cílem této kapitoly je odpovědět na otázku, proč by se v oblasti SLA[2] mělo (více) počítat s frekvencí. Postupovat v ní budu od obecného shrnutí mezinárodní diskuze o roli, jakou může frekvence v lingvistice sehrávat (2.1), k bližším zastavením u přístupů, s jejichž východisky se ztotožňuje tato práce (2.2–2.4).

V českém prostředí pojem frekvence tradičně nepřekračuje oblast korpusové, matematické či kvantitativní lingvistiky,[3] v těchto rámcích ovšem v současnosti výrazně nabývá na významu jako korpusový ukazatel povahy, potažmo důležitosti určitého jevu (NESČ). Dokladem těchto tvrzení může být například skutečnost, že *Nový encyklopedický slovník češtiny* (v tisku), na rozdíl od staršího *Encyklopedického slovníku* (2002), zpracovává frekvenci jako samostatné heslo, i to, že autorem hesla je Václav Cvrček.[4] Cvrček a kol. již na frekvenčních údajích postavili rozsáhlý popis současného jazyka – dvoudílnou *Mluvnici současné češtiny.* Frekvenci přitom důvěřují jako ukazateli toho, „jak [jazyk] skutečně vypadá" (MSČ: 13). Na jeho základě je různým jevům, formám a variantám možné připsat „takovou váhu, jaká jim reálně v úzu náleží" (tamtéž).

Zahraniční lingvistika jde v přisuzování důležitosti frekvenci ještě o něco dál. Ačkoli pohled na ni není ustálen a v diskuzích o limitech jejího vlivu na jazyk a jeho mluvčí se objevují i názory, že frekvence není vhodným konceptem pro explanaci jazykových jevů nebo že je jen

[2] Osvojování druhého jazyka (Second Language Acquisition).

[3] Jako počátek systematického odborného zájmu o frekvenční charakteristiky českého lexikonu lze chápat práce kvantitativní lingvistky Marie Těšitelové z osmdesátých let minulého století.

[4] Význam frekvence v rámci české korpusové lingvistiky charakterizuje Cvrček tato: „Frekvence jako základní veličina libovolné jednotky (typu) a languová (systémová) charakteristika se používá nejen k poměřování mezi alternujícími jevy (např. frekvence morfologických variant bychom a bysme, viz SyD), ale slouží také ke konstruování slovníků (vymezení nejčetnějších slov jako jádra slovní zásoby), extrakci kolokací, zhodnocení gramatických kategorií, identifikaci klíčových slov v textech apod." (Cvrček – Richterová 2013; v zásadě táž definice se objevuje i v NESČ).

epifenoménem jevů jiných (Roeper 2007),[5] stala se jedním ze základních konceptů funkčních a kognitivních přístupů k jazyku, nebo se alespoň diskuze o ní odehrává právě v tomto rámci.

Dagmar Divjak a Catherine Caldwell-Harris (2015) v nejnovějším kompendiu kognitivní lingvistiky uvádějí, že v psycholingvistice a kognitivní lingvistice se frekvencí obvykle rozumí počet případů, kolikrát se určitý jazykový prvek (foném, morfém, slovo nebo fráze) objeví v určitém prostředí, a že se obvykle užívá v relativním smyslu (aby se určitý jev dal zhodnotit jako více či méně běžný) (Divjak – Caldwell-Harris 2015: 54). Touto základní definicí se frekvence v kognitivním rámci ještě nijak zásadně neliší od definice čistě korpusové (srov. 4). Aspekt, kterým se již liší, je předpoklad korelace mezi frekvencí určitého jevu a jeho zakotvením v mysli.[6] Předpoklad, že čím častěji je mluvčí určitému jazykovému jevu vystaven, tím silněji je daný jev zakotven v mysli, je již několik desetiletí předmětem psycholingvistického výzkumu, a psycholingvistické bádání tento předpoklad do značné míry potvrzuje. Ukazuje se, že zpracování jazyka je mimořádně citlivé k frekvenci užívaní, a to na všech rovinách lingvistické reprezentace – „[l]anguage knowledge involves statistical knowledge, so humans learn more easily and process more fluently high frequency forms and ‚regular' patterns which are exemplified by many types and which have few competitors" (Ellis 2014: 196; pro podrobnější přehled vlivu frekvence prvků z různých jazykových rovin na jejich zpracování viz Ellis 2002). Jazykový systém rodilého či nerodilého mluvčího se pak dá chápat jako výsledek zpracování statistických informací.

[5] Tom Roeper (2007) mj. kritizuje přímé spojování frekvence s mentální reprezentací – numeralizuje se jím něco, čemu mechanismus počítání není vlastní. K objasnění problému používá Roeper příměr k opotřebovaným botám. Boty, které urazí 1000 kroků, budou opotřebovanější než boty, které jich ujdou jen 100, počet kroků tedy může soužit k predikci opotřebovanosti obuvi, neznamená to ale, že by samy boty kroky počítaly. Frekvence je nepřímou metodou, u které je nezbytné zhodnotit, zda je spojena s jádrovým mechanismem, nebo jen od něj jen odvozena, tak jak je tomu podle Roepera v případě kroků i jazykových jednotek. S pojmem frekvence je podle Roepera korektní operovat pouze v případě, kdy se k mentální reprezentaci nepřidává žádná nová informace. Jeho výtky vůči frekvenci jako explanatornímu konceptu ovšem pochází z oblasti osvojování prvního jazyka a pro SLA mají jen omezenou platnost.

[6] Pojem zakotvení (*entrenchment*) zavedl Ronald Langacker a předpokládal pro něj ve vztahu k užívání určité struktury následující tendence: "Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence." (Langacker 1987: 59, cit. dle Divjak – Caldwell-Harris 2015: 60)

2.1.1 A s čím dalším je třeba počítat?

Gass a Mackey v roce 2002 jako vysoce důležité téma, kterým by se měl zabývat budoucí výzkum, hodnotí problematiku interakce frekvence s jinými faktory procesu osvojování druhého jazyka. Na další faktory, které v úspěšnosti osvojení určité struktury mohou hrát roli, je zpravidla upozorňováno i v současné diskuzi o rozsahu vlivu frekvence. Například Ellis (2014: 195) jako faktory tradičně vymezované v psychologickém výzkumu uvádí (vedle frekvence) recenci/recentnost a kontext – „the more times we experience something, the stronger our memory of it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory of it, and the more fluently it is accessed. The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization; so a stimulus becomes associated to a context and we become more likely to perceive it in that context." Sám pak faktory ovlivňující SLA třídí do čtyř oblastí, které se týkají (a) frekvence inputu, (b) formy, (c) funkce a (d) interakce mezi formou a funkcí (Ellis 2014: 198nn).

Mezi faktory (a) Ellis vedle frekvenčních efektů, kterým se budu podrobněji věnovat v následujících oddílech, řadí právě výše zmíněnou recenci. Označuje jí efekt známý jako priming, tedy vyvolání určité struktury jinou/podobnou strukturou, které je mluvčí vystaven. Do oblasti (b) podle Ellise spadá salience a percepce. Jde o míru důležitosti/významnosti, kterou mluvčí připisuje vnímanému podnětu, resp. míru nápadnosti daného podnětu. Ellis (2014: 200) například uvádí, že z prostředků odkazujících k přítomnému času je více salientní lexikální jednotka *today* než morfém pro třetí osobu singuláru *-s*, *today* tedy pravděpodobněji upoutá pozornost mluvčího. V oblasti (c) věnuje Ellis pozornost prototypičnosti významu a redundanci. Prototypem, tj. nejlepším reprezentantem určité kategorie, nutně nemusí být jednotka o největší frekvenci, ale platí, že čím vyšší je tokenová frekvence příkladu, tím lépe přispívá k vymezení určité kategorie a tím vyšší je pravděpodobnost, že bude pokládán za prototypický (Ellis 2014: 201). Pokud jde o redundanci, platí, že jednotky, které jsou při zpracování jazykového podnětu nadbytečné (například vyskytují-li se s jinými jednotami, které zprostředkují týž význam snadněji), mají tendenci „nebýt osvojené" (tamtéž). Jako příklad Ellis opět uvádí nepotřebnosti morfologického značení času v konkurenci s užitím časového adverbiálního výrazu.

Existuje tedy řada dalších, empiricky ověřených faktorů, se kterými je třeba vedle frekvence počítat. Divjak a Caldwell-Harris (2015: 68) explicitně zodpovídají otázku, zda je frekvence

nejdůležitějším faktorem pro zakotvení reprezentace v mysli, záporně. Zdůrazňují především váhu kontextu a mimo to jako faktor, který může přebít frekvenční efekty, zmiňují relevanci podnětu pro konkrétního mluvčího. Pokud je například nějaký jazykový podnět spojen se silným emocionálním prožitkem, může být na základě i jediného výskytu utvořena jeho silná mentální reprezentace (Divjak a Caldwell-Harris 2015: 69).[7]

2.2 Přístupy založené na užívání

Význam frekvence pro kognitivní procesy zdůrazňuje ve svých pracích především Joan Bybee, průkopnice teorií založených na užívání (*usage-based*).[8] Přístupy založené na užívání předpokládají, že frekvence podnětu, kterému je mluvčí vystaven, má vliv na úspěšnost jeho osvojení a zakotvení v mysli (Ellis – Wulff 2015: 410; Bybee: 2008), a na gramatiku nahlížejí jako na kognitivně organizované zkušenosti, které mají mluvčí s jazykem (Bybee 2006; 2008). „Zažité" jazykové prvky jsou v mysli organizovány jako rozsáhlá sít fonologických, sémantických a pragmatických spojů překračující dělení na gramatiku a lexikon, utvářená opakováním.

Ve studii vztahující přístupy založené na užívání k teoriím osvojování druhého jazyka hodnotí Bybee význam frekvence pro úspěšné nabytí jinojazyčného kódu jako v určitém ohledu špatnou zprávu pro nerodilého mluvčího (Bybee 2008: 232).[9] Student, který druhý/cizí nabývá v dospělém věku nebo ve výukovém prostředí, nebude cílovému jazyku nikdy vystaven jako rodilý mluvčí. Dobrou zprávou je, že přirozená frekvenční distribuce alespoň těch nejčetnějších jednotek je vyšší, než je k osvojení třeba (Bybee 2008: 233).

2.3 Frekvenční efekty

Pro uchopení povahy frekvence v rámci přístupů založených na užívání i mimo ně je stěžejní distinkce mezi frekvencí tokenovou (*token frequency*) a frekvencí typovou (*type frequency*). Toto rozlišení umožňuje Joan Bybee demonstrovat vliv, který má opakování určité jazykové

[7] Emocionální prožitkem může být např. očekávání nebo překvapení (viz Ellis 2006).

[8] Nejsem si vědoma toho, že by v češtině pro *usage-based* přístupy existoval ustálený ekvivalent. Karin Schöne, která *usage-based* přístupy zmiňuje ve své disertační práci, např. volí termín přístupy *založené na úzu*. Takový překlad nepovažuji za zcela přesný. V této práci budu používat termín přístupy *založené na užívání*.

[9] Nadto představuje opakování pro nerodilého mluvčího možné nebezpečí v podobě fosilizace (Bybee 2008: 221).

jednotky na její kognitivní reprezentaci (Bybee 2008: 218). Oba druhy frekvencí mají vliv na úspěšnost osvojování druhého jazyka.

V tomto oddíle na základě prací Bybee a kol. vymezím rozdíl mezi tokenovou a typovou frekvencí a na vlastních, pro téma práce relevantních příkladech demonstruji frekvenční efekty,[10] které se na tomto rozlišení zakládají (Bybee 2008, Bybee 2010).

2.3.1 Tokenová a typová frekvence

Tokenová frekvence se vztahuje k počtu výskytů určité jednotky v textu. Přičemž onu jednotku lze chápat v zásadě na libovolné úrovni komplexnosti – lze zjišťovat tokenovou frekvenci konsonantu, slabiky, slova, fráze i věty. Při zkoumání substantivní morfologie tak například můžeme určit počet všech výskytů substantiv v lokálu a v instrumentálu v korpusu CZESL (tj. 24 135 tokenů pro lokál a 10 794 pro instrumentál).

Typová frekvence se vztahuje k výskytu určitého vzorce (*pattern*). Jednoduše řečeno se jedná o počet rozdílných jednotek (bez opakování), které daný vzorec reprezentují. Za typovou frekvenci by se tedy například, analogicky k příkladu prvnímu, dal pokládat počet všech různých substantiv, bez ohledu na jejich opakování, které v korpusu CzeSL stojí v lokálu a v instrumentálu (tj. 2 008 typů pro lokál a 2 008 pro instrumentál).

2.3.2 Konzervační efekt

Prvním ze tří efektů, které Bybee připisuje tokenové frekvenci, je tzv. konzervační efekt (*conserving effect*). Tento efekt vychází z kognitivního předpokladu, že opakování určité jednotky posiluje její reprezentaci v paměti, činí ji odolnější a snáze přístupnou (*accessible*). Konzervační efekt tak vysvětluje, proč nepravidelné jednoty s vysokou frekvencí nepodléhají vyrovnávání na základě analogie s pravidelnými formami, zatímco méně frekventované ano. Jako příklad uvádí Bybee tendenci, pozorovatelnou jak v dlouhodobém vývoji angličtiny, tak vývoji dětské řeči: Nepravidelná slovesa s nízkou frekvencí podléhají v minulém čase vlivu pravidelné konjugace (původně *wept*, *crept*, dnes *weeped*, *creeped*), zatímco vysoce frekventovaná slova stejné třídy si nepravidelná tvary minulého času zachovávají (*kept*, *slept*).

Pro českou substantivní morfologii lze například předpokládat, že alternace hlásky *h* v hlásku *z* v dativu či lokálu singuláru bude nerodilému mluvčímu činit menší potíže v případě slova *noha* (33 693 tokenů v korpusu SYN2015), než u slova *duha* (509 tokenů tamtéž), kde se bude

[10] Srov. Křivan (2012), který efekty vymezované Bybee demonstruje na českém stupňování.

více projevovat tendence k zachování *h* analogicky k ostatním pádům. Výsledky korpusu nerodilých mluvčích (CzeSL-SGT) tento předpoklad potvrzují, byť na minimálním vzorku.[11] V požadovaných tvarech se v korpusu *noha* a *duha* objevují takto:

(1) udělat obvaz. Následující den ještě mě bolelo v < noze >. Jeli jsme k jinému lekaři a ten řekl mi že mám

(2) toho vůbec nevšimli že spí. protože stali na jedné < noze >. Když jsem otevřela jejich místnost, začali křičet

(3) Jakou barvu má život; Život člověka je podoben < duhě > . Každý život každého člověka má bilou a černou část

Ačkoli korpus celkem obsahuje pouze tři výskyty, i na nich lze demonstroval konzervační efekt.

2.3.3 Efekt autonomie

Jako specifický efekt Bybee dále vyčleňuje autonomii (*autonomy*), která může být chápána jako extrémní případ konzervačního efektu. Čím autonomnější je určitá jednotka, tím větší je pravděpodobnost, že bude v mysli reprezentovaná zvlášť, odděleně od jednotek ze stejného paradigmatu, bez asociací k jiným jednotkám. Analogicky k příkladu, který uvádí Bybee se svou zkušeností se španělštinou (2008: 219), by bylo možné uvažovat o tom, že zatímco slovo *dovolená* tvoří v akuzativu singuláru součást paradigmatu *dovolené*, slovotvorně shodná *shledaná*, u níž by se nízké povědomí o příslušnosti k nějakému paradigmatu pravděpodobně ukázalo i u rodilých mluvčích, bude reprezentována autonomně. Usouvztažněno s chybami v korpusu CzeSL, ve tvaru (*na*) *dovolenou* je chybná koncovka užita v 74 případech 314, ve tvaru (*na*) shledanou v jednom případě ze 46, což ukazuje minimálně to, že roli v úspěšnosti osvojení určité formy nehraje jen čistá frekvence.

2.3.4 Redukční efekt

Poslední efekt, připisovaný tokenové frekvenci, je redukční efekt. Redukční efekt se týká běžně pozorovatelného jevu, a to fonetické redukce. Více frekventované jednotky podléhají častěji fonetické redukci než jednotky méně frekventované.

[11] Příklady mají daný efekt ilustrovat, nikoli dokazovat.

2.3.5 Typový efekt

Typová frekvence je hlavní faktor určující stupeň produktivity konstrukce. Pokud je určitá konstrukce aplikovaná na vysoký počet různých jednotek, bude pravděpodobně aplikovatelná i při tvoření nových jednotek. Vedle frekvence jsou pro produktivitu konstrukce ovšem podstatné též jiné faktory, jako např. sémantická a morfologická složitost.

„Čím více existuje tvarů, které nesou daný afix, tím je silnější reprezentace tohoto afixu. Čím silnější je reprezentace afixu, tím snadnější k němu bude přístup, když bude nové slovo vyžadovat flexi, a tím větší bude pravděpodobnost, že tento afix bude produktivní." (Bybee 1995, cit. dle Křivan 2012: 17)

To pro osvojování morfologie může například znamenat, že v rámci deklinačního vzoru hrad budou mít nerodilí mluvčí tendenci upřednostňovat v genitivu singuláru koncovku -u před koncovkou -a, neboť se jedná o tvar, který tvoří na 90 % substantiv vzoru hrad, a je tedy silněji reprezentován.[12]

[12] Údaje o frekvenčním zastoupení těchto tvarů jsou čerpány z MSČ. Srovnatelný předpoklad formuluje Schöne (2015: 17) pod termínem zásada dominantního paradigmatu (vycházejíc ze zásady *dominant paradigm condition* formulované Stephanem Schmidem (1997)).

2.4 Gramatické profily

Použití konceptu gramatických profilů v této práci jako nástroje explanace osvojování české substantivní morfologie nerodilými mluvčími vychází především z prací kognitivní lingvistky Laury Jandy a kol. (Janda – Lyashevskaya 2011, Eckhoff – Janda 2014) a Stefana Griese a Dagmar Divjak (Divjak – Gries 2006, 2008; Gries – Divjak 2009).

2.4.1 Behaviorální profily

Gries a Divjak (2009) ve své studii, zabývající se dvěma ústředními významovými vztahy, polysémií a synonymií, a otázkou po možnostech určovaní podobností a vztahů mezi významy či slovy, kritizují dosavadní pokusy o uchopení této problematiky pro nedostatečnou empiričnost. Popularita termínu *usage-based* se podle nich projevuje spíše v kognitivně lingvistických teoriích než v praxi, která nedostatečně využívá potenciálu korpusové lingvistiky (Gries a Divjak, 2009:59). Sami proto navrhují přístup radikálně založený na korpusu (*radically corpus-based*). Jako radikální ho označují proto, že mnohem více než předchozí studie vychází z předpokladu, že mezi distribučními vzorci a funkčními vlastnosti existuje korelace,[13] tedy například že podobné distribuční vlastnosti indikují podobné vlastnosti funkční. Cílem je propojení korpusové a kognitivní lingvistiky – uplatnění důsledné, kvantitativní analýzy založené na korpusu, které povede ke kognitivně relevantním výsledkům (2009: 60).
Korpusová analýza Griese a Divjak se zakládá na dvou konceptech – a) tzv. ID tagech a b) behaviorálních profilech:[14]

a) ID tagy označují všechny relevantní informace, které lze o určitém slovesu[15] získat z větného kontextu, v němž se dané sloveso vyskytuje – formální charakteristiky slovesa, věty, sémantické vlastnosti slovesa a jiných členů věty atp. (Divjak – Gries

[13] Funkční vlastnosti jsou zde míněny v širokém slova smyslu, zahrnují sémantické, pragmatické, diskurzivní aj. funkce (Gries a Divjak, 2009: 59).

[14] Oba pojmy už jsou v lingvistickém oběhu delší dobu. V případě konceptu ID tagů vycházejí Gries a Divjak ze studie Sue Atkinson (1987), v případě behaviorálních profilů ze studie Patricka Hankse (1996). Pro účely této práce jejich vymezení přejímám od Griese a Divjak (2009).

[15] Gries a Divjak (2009; Divjak – Gries 2006), stejně jako Janda a kol. (Janda – Lyashevskaya 2011; Eckhoff – Janda 2014) aplikují ve svých studiích pojem behaviorálních či gramatických profilů pouze na analýzu sloves, nikterak však vztažení pojmu na jiný slovní druh nevylučují.

2006: 28). Pro přesnější představu cituji tabulku, ve které Gries a Divjak (2009) shrnují používané ID tagy s dodatkem, že se nejedná o výčet vyčerpávající a že může být doplněn dalšími ID tagy nebo druhy ID tagů – např. z pragmatické nebo fonologické domény.

Kind of ID tag	ID tag	Levels of ID tag
morphological	tense	present, past, future
	mode	infinitive, indicative, subjunctive, imperative, participle, gerund
	aspect	imperfective vs. perfective
	voice	active vs. passive
	number	singular vs. plural
	transitivity	intransitive, monotransitive, copular, complex transitive
syntactic	sentence type	declarative, exclamative, imperative, interrogative
	clause type	main vs. dependent
	type of dependent clause	adverbial, appositive, relative, zero-relative, zero-subordinator, etc.
semantic	semantic types of subjects, objects, etc.	concrete vs. abstract, animate (human, animal) vs. inanimate (event, phenomenon of nature, body part, organization/institution, speech/text) etc.
	countability of nouns	count vs. mass
	properties of the process denoted by the verb	physical actions, physical perception, communication, intellectual activities, emotions, etc.
	controllability of actions	high vs. medium vs. no controllability
	adverbial/PP modification (if present)	PP temporal, locative, etc.
	negation	present vs. absent, attached to which element
lexical	collocates in precisely defined syntactic slots (i.e. collexemes)	$collocate_1$, $collocate_2$, …, $collocate_n$

Tabulka 2.1 Přehled (druhů) ID tagů a jejich úrovní (Gries a Divjak 2009:62)

b) Behaviorální profily vznikají na základě výše uvedených ID tagů. Přesněji řečeno jsou behaviorální profily určitých sloves nebo jednotlivých významů sloves dány souhrnem hodnot jejich ID tagů, resp. frekvenční distribucí těchto hodnot.

Výhodou behaviorálních profilů založených na frekvenční distribuci jsou široké možnosti jejich vyhodnocení pomocí statistických metod. Divjak a Gries (2009) představují využití monofaktoriální a multifaktoriální metody analýzy.

2.4.2 Gramatické profily

Pojem gramatické profily používá ve svých studiích Laura Janda a kol. (Janda – Lyashevskaya 2011; Eckhoff – Janda 2014). Gramatické profily vychází z behaviorálních profilů Divjak a Griese (podle autorek by se daly popsat jako jejich podtyp), nezahrnují ovšem tak komplexní množinu rysů, jako postihují ID tagy. Omezují se pouze na morfologické vlastnosti zkoumaných jednotek, což je motivováno snahou o snadnější manipulaci s daty (Eckhoff – Janda 2014:233) a pozorováním, že některá slovesa jsou v některých tvarech užita mnohem častěji než jiná (Janda – Lyashevskaya 2011: 720).[16] Gramatické profily jsou tedy tvořeny informací o frekvenční distribuci určitých sloves.

Ve zmiňovaných studiích přispívají gramatické profily sloves k výzkumu otázek spjatých s videm v současné ruštině (Janda – Lyashevskaya 2011) a ve staroslověnštině (Eckhoff – Janda 2014). V obou případech se prokázalo, že statistická analýza gramatických profilů sloves umožnuje klastrování sloves do tříd, které vykazují relevantní vidové rozdíly.

2.4.3 Morfologické profily

Termín morfologické profily užívá ve své disertační práci a s ní souvisejícím článku Karin Schöne (2015, 2011). Schöne se morfologickými profily zabývá v souvislosti s výukou češtiny jako cizího jazyka. Jde jí o to, „zjišťovat a analyzovat frekvenční distribuci substantivních pádových tvarů v současné češtině“ (2015: 8). Současnou češtinou se zde rozumí čeština synchronního korpusu psaných textů k roku 2005 (SYN2005 ÚČNK) a mluveného korpusu ORAL2006. Práce má v současných psaných i mluvených projevech rodilých mluvčích identifikovat typická užití zkoumaných lemmat.[17] Schöne ukazuje, že je možné substantiva

[16] Samostatný problém představuje otázka míry podrobnosti, kterou by analýza korpusových dat měla zohledňovat (viz např. Janda – Lyashevskaya 2011; Gries 2011). Zhodnocení zahraniční diskuze o tom, zda analýza na obecnější rovině (lemmat) neposkytuje přesnější výsledky a představu o zkoumaném jevu než analýza na rovině podrobnější (slovních tvarů), stojí mimo rámec této práce. Mimo to v případě tak flektivního jazyka, jako je čeština, a substantivní, nikoli verbální flexe relevance této problematiky poněkud slábne.

[17] K výzkumu jsou vybírána taková lemmata, která by měl nerodilý mluvčí češtiny ovládat na úrovni B1. Toto množina je sestavována jednak na základě Popisu referenční úrovně B1, jednak s přihlédnutím k frekvenčnímu slovníku češtiny. Vzorek čítá 203 slov.

třídit do specifických skupin na základě dvou (v některých případech pro podrobnější klasifikaci tří) nejfrekventovanějších tvarů. K vymezeným třídám budu přihlížet v analytické části.

V této práci se gramatickými profily rozumí substantivní třídy/klastry lemmat se stejnou frekvencí, které byly vyčleněné na základě tří či (kvůli limitům malého vzorku) dvou nejfrekventovanějších pádových tvarů jednotlivých substantivních lemmat.

3 Metodika výzkumu

3.1 Korpusy nerodilých mluvčích

V současné době existují pro češtinu dva (na poměry akvizičních korpusů) rozsáhlé a volně dostupné korpusy písemných projevů nerodilých mluvčích – CzeSL a Merlin.[18] Při své analýze budu zohledňovat data získaná z obou těchto korpusů. CzeSL vzhledem ke svému podstatně většímu rozsahu (téměř jeden milion slov) oproti korpusu Merlin (několik desetitisíců slov) nabízí možnost zkoumat gramatické profily substantiv,[19] Merlin zase díky přesnější (ruční) anotaci chyb umožňuje např. zasadit chyby, které v korpusu CzeSL kvůli časové náročnosti vlastní manuální anotace budu pozorovat jen na několika příkladech nejfrekventovanějších lemmat, do širšího kontextu lemmat dalších. V tomto oddíle nejprve stručně charakterizuji oba korpusy, v oddílu dalším (3.2) popíšu, jak jsem postupovala při sestavování vzorků z těchto korpusů a jeho ruční anotaci.

3.1.1 Merlin

Mezinárodní projekt Merlin představuje trojjazyčný korpus obsahující texty nerodilých mluvčích češtiny, němčiny a italštiny, který je budován primárně za účelem poskytnout uživatelům dostatek autentického materiálu, jenž by dokládal různé úrovně jazykové kompetence tak, jak je vymezuje *Společný evropský referenční rámce pro jazyky* (SERR).

[18] Korpus CzeSL je dostupný z: https://kontext.korpus.cz; Merlin je dostupný z: http://merlin-platform.eu.

[19] Paradigma určitých lemmat je v něm zastoupeno takovým počtem tvarů, aby v pořadí frekvencí těchto tvarů bylo možné sledovat určité tendence.

„Merlin reaguje na poptávku po názorných příkladech úrovní SERR […].“[20] Jak se lze dočíst na jeho oficiálních stránkách, hlavním cílem projektu není přispět svými daty k rozvoji výzkumu cizího jazyka, nýbrž jeho didaktiky, a to ve čtyřech oblastech – a) při jazykové výuce, b) při tvorbě učebních matriálů, c) při vytváření sylabů a kurikul, d) při tvorbě testů. Nabízí tedy potenciál pro přímé i nepřímé využití korpusu pro výuku češtiny (němčiny a italštiny) jako cizího jazyka.[21]

Na české straně se na projektu podílí Ústav odborné a jazykové přípravy UK a jeho řešitelé v současné době začínají publikovat články představující korpus v českém prostředí (Štindlová – Čurdová 2015, Pečený 2015, Štindlová a kol. 2014). Korpus je tvořen texty vzešlými z mezinárodních zkoušek, českou část tvoří texty z Certifikované zkoušky z češtiny pro cizince (CCE), jejichž vytváření ve shodě se SERR má v kompetenci právě ÚJOP UK.

Celkově korpus sestává z 2 286 textů, z toho 441 českých, které byly vytvořeny v rámci zkoušek pro úrovně A2, B1 a B2, což celkem pro češtinu odpovídá 79 969 tokenům.

Vícejazyčný charakter projektu a jeho spojení se SERR klade specifické nároky na pravidla anotace, která musí být stejným způsobem aplikovatelná na různé evropské jazyky – slovanské, germánské i románské – a která má postihovat jejich společné rysy i jejich specifika.

Při anotaci textů jsou v projektu Merlin stanoveny dvě cílové hypotézy, tj. rekonstrukce/interpretace projevu žáka na dvou úrovních. První cílová hypotéza (TH1, *minimal target hypothesis*) se vztahuje k úrovni ortografické a gramatické, snaží se o minimální počet zásahů do pravopisu, morfologie a syntaxe, jejichž výsledkem je gramaticky správně zformovaná věta, měla by zůstat věrná povrchové realizaci projevu studenta, jak je to jen možné. Druhá cílová hypotéza (TH2, *extended target hypothesis*) se týká rovin vyšších, lze na ní upravovat lexikální, sémantické a stylistické aspekty textu a více zohlednit pravděpodobnou

[20] Z popisu projektu na stránkách ÚJOP UK, dostupné z: http://ujop.cuni.cz/merlin-program-celozivotniho-vzdelavani. Kritiku SERR a jeho nedostatečné empirické základny shrnuje např. Schöne (2015).

[21] Zároveň obsahuje platforma Merlin řadu didaktických tipů a manuálů (a videomanuálů). V oblasti a) například navrhuje poměrně náročný způsob využití: „Přeneste platformu MERLIN do vaší třídy: Můžete nechat své (pokročilé) studenty vyhledávat jazykové jevy v subkorpusu MERLIN, aby se s tímto nástrojem seznámili a zároveň podpořili svou samostatnost v učení se cizího jazyka. Mohou společně analyzovat jak příklady chyb nalezené v databázi MERLIN, tak své vlastní texty. Můžete je také nechat porovnat výsledky vyhledávání v platformě MERLIN s národním korpusem, aby tak získali představu o rozdílech v užívání různých jazyků.“

intenci autora. TH1 se tedy zaměřuje na jazykovou správnost, TH2 na vhodnost/přiměřemost vyjádření (Boyd a kol. 2014: 1284).

Anotační schéma je tříúrovňové. První rovina vymezuje kategorii (např. gramatika), druhá její podtyp, tzv. rys (např. chybné flexe). Tyto roviny jsou obligatorní. Na třetí rovině se v některých případech uvádí specifikace odchylky (Boyd a kol. 2014: 1284; Štindlová a kol. 2014: 143n).

Jednou z problematických oblastní anotace, zvláště pro češtinu, je rozhodování o chybách na pomezí ortografie a morfologie. Explicitně se jí zabývají i Štindlová a kol. (2014: 146) a Štindlová a Čurdová (2015: 198). Hodnocení jakékoli chyby v koncovce flektivních tvarů jako gramatické, jak předepisuje anotačních schéma CzeSLu, považují za příliš zjednodušující a pro češtinu v mnoha případech zcela neintuitivní. Štindlová a kol. (2014) proto pro rozhodování o těchto případech vytvořili následující algoritmus, kterým bylo anotační schéma Merlinu doplněno:

<table>
<tr><td rowspan="5">Jedná se o chybu v tvaroslovné charakteristice (TCH)?</td><td rowspan="3">ANO</td><td rowspan="3">Existuje v paradigmatech daného slovního druhu?</td><td>ANO</td><td colspan="3">všichni jsou spokojené</td><td>→ chyba v morfologických kategoriích</td></tr>
<tr><td rowspan="2">NE</td><td rowspan="2">Existuje v paradigmatech daného slovního druhu?</td><td>ANO</td><td>večeřit</td><td>→ chyba ve flexi</td></tr>
<tr><td>NE</td><td>kratke kalhoty</td><td>→ některá z ortografických chyb</td></tr>
<tr><td rowspan="2">NE</td><td rowspan="2">Slovotvorná motivace chyby?</td><td>ANO</td><td colspan="3">němečtina</td><td>→ slovotvorná chyba</td></tr>
<tr><td>NE</td><td colspan="3">tatinek</td><td>→ některá z ortografických chyb</td></tr>
</table>

Tabulka 3.1 Anotační algoritmus (Štindlová a kol. 2014)

Pokud se tedy chybný tvar shoduje s jiným tvarem paradigmatu, je mu připsána chyba gramatická, pokud se takový tvar v paradigmatu nenachází, jedná se o chybu pravopisnou. Pokud tak například nerodilý mluvčí bude psát v singuláru lokálu o *zemí* a o *muží*, bude se v případě slova *země* jednat o chybu gramatickou, v případě *muže* o pravopisnou. Rozhodování o klasifikaci tohoto druhu chyby bude ještě předmětem diskuze v oddíle 3.3.1.

3.1.2 CzeSL

Korpus CzeSL (Czech as a Second Language) je dlouhodobě vyvíjen ve spolupráci Technické univerzity v Liberci, Univerzity Karlovy a Asociace učitelů češtiny jako cizího jazyka.[22] Co do počtu textů nerodilých mluvčích se řadí k nejrozsáhlejším žákovským (neanglických) korpusům (Šebesta 2012: 29).

V současné době jsou zveřejněny jako dvě jeho verze – CzeSL-plain a CzeSL-SGT.[23]
CzeSL-plain je prvním výstupem projektu a zahrnuje přibližně 2,3 milionů tokenů bez jakékoli lingvistické anotace. Zastřešuje tři subkorpusy různých typů (Cvrček – Richterová 2015):

a) ciz – přepisy písemných prací (esejů) nerodilých mluvčích, které vznikly v souvislosti s jazykovým vyučováním v kurzech různého druhu a úrovně;
b) kval – odborné texty získané od nerodilých mluvčích studujících na českých vysokých školách v navazujícím magisterském či doktorském studiu;
c) rom – přepisy školních písemných prací romských žáků z oblastí ohrožených sociálním vyloučením.

Subkorpus (a) ciz obsahuje celkem 8 109 textů. Na tomto základě vznikl korpus CzeSL-SGT, který tuto část korpusu CzeSL-plain obsahující eseje cizinců z let 2009–2012 doplňuje ještě o texty získané v roce 2013, čímž dosahuje velikosti celkem 960 000 slov.[24]

CzeSL-SGT (*Czech as a Second Language with Spelling, Grammar and Tags*) se od korpusu CzeSL-plain zásadně odlišuje tím, že je lemmatizován, vybaven standardní anotací pro slovní

[22] Jako jeden z výstupů projektu *Inovace vzdělávání v oboru čeština jako druhý jazyk* v rámci OP Vzdělávání pro konkurenceschopnost, s finanční podporou Strukturálních fondů EU (ESF) a státního rozpočtu České republiky.

[23] Obecně korpus CzeSL budí dojem, že se studie píšou spíše o něm než na základě něho. Může to dokládat například rešerše v repozitáři závěrečných prací, které vznikají na FF UK – z celkem patnácti prací, které slovo CzeSL obsahují, se ve většině případů jedná o pouhou zmínku o existenci tohoto korpusu (konstatování, že se připravuje a že je jedná o unikátní projekt), v dalších několika případech se využívá jeho chybová anotace (jako inspirace k vlastní anotaci i jako materiál pro studii možností chybových anotací), jen v jediném případě jsou využívány texty tohoto korpusu k analýze. Jako další doklad výše uvedeného tvrzení může být fakt, že někdy se zveřejněný korpus CZeSL (popř. jeho verze CZeSL-SGT) popisuje jako korpus s manuální anotací (např. Schöne 2015). Ke zveřejnění ručně anotované verze CZeSL ovšem ještě nedošlo a verze, které jsou dostupné, jsou anotované pouze automaticky.

[24] Tento údaj je uveden na stránkách Ústavu českého národního korpusu (http://ucnk.ff.cuni.cz), Rosen (2015) uvádí 1 148 000 tokenů.

druh a morfologické kategorie a chybovou anotací. Všechny tyto procesy byly provedeny automaticky. Vzhledem k chybové anotaci se tradiční atributy (word, lemma a tag) dvojí. Word značí původní slovní tvar, lemma určuje reprezentativní tvar původního slovního tvaru a tag slovnědruhové a morfologické značky původního tvaru. K těmto atributům se přivává word1, lemma1 a tag1 značící totéž pro tvar opravený. Dále je každý chybný tvar označen atributem gs, který určuje, zda se jedná o chybu pravopisnou (S) či gramatickou (G), a atributem err, který na základě porovnání původního a opraveného tvaru blíže určuje typ chyby (Rosen 2015).[25]

Z dokumentace korpusu CzeSL-SGT je důležité zmínit ještě dva momenty, které se ukážou určující pro postup tohoto výzkumu (viz oddíl 3.3). První se týká postupu automatické lemmatizace: „pokud tvar není rozpoznán, je lemma totožné s původním tvarem", druhý automatického vyhodnocování chyby a atributu gs: „gramatická chyba se obvykle vyznačuje tím, že původní slovní tvar byl rozpoznán“ (Rosen 2015).

3.2 Stavba vzorku

Pro analýzu chybovosti v substantivních pádových tvarech byla vybrána slova, která v synchronním korpusu češtiny SYN2015[26] patří mezi nejfrekventovanější. SYN2015 je reprezentativní a referenční korpus psaného jazyka čítající přes 100 milionů tokenů (Cvrček – Richterová 2015). Následující tabulka shrnuje 200 jeho nejfrekventovanějších lemmat.

[25] Kompletní seznam použitých chybových značek je dostupný na: http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1.html.

[26] Dostupný z: https://kontext.korpus.cz/first_form.

r	g	p	lemma	f	r	g	p	lemma	f	r	g	p	lemma	f	r	g	p	lemma	f
1.	I	1	rok	358 356	51.	M	3	otec	38 344	101.	N	12	světlo	26 657	151.	I	57	model	21 299
2.	M	1	člověk	203 726	52.	F	28	rodina	38 284	102.	M	5	pan	26 478	152.	N	20	srdce	21 190
3.	I	2	den	111 726	53.	I	16	měsíc	38 120	103.	I	39	zápas	26 305	153.	M	12	autor	21 065
4.	F	1	doba	108 001	54.	F	29	matka	37 123	104.	N	13	prostředí	26 011	154.	M	13	Petr	21 038
5.	N/F	28	dítě	104 530	55.	F	30	změna	36 652	105.	F	47	válka	25 895	155.	F	63	láska	21 030
6.	I	3	život	96 721	56.	I	17	týden	36 575	106.	I	40	cíl	25 823	156.	N	21	dílo	20 977
7.	N	1	místo	95 819	57.	N	6	jméno	36 437	107.	I	41	pokoj	25 499	157.	F	64	polovina	20 950
8.	F	2	ruka	89 653	58.	I	18	projekt	36 406	108.	M	6	rodič	25 245	158.	F	65	pomoc	20 944
9.	F	3	práce	81 296	59.	N	7	auto	35 836	109.	F	48	osoba	25 033	159.	I	58	strom	20 866
10.	I	4	svět	79 949	60.	F	31	tvář	35 529	110.	N	14	řešení	24 959	160.	I	59	příběh	20 798
11.	F	4	strana	79 801	61.	I	19	výsledek	35 068	111.	F	49	barva	24 857	161.	F	66	stránka	20 793
12.	F	5	země	77 054	62.	F	32	většina	34 973	112.	F	50	funkce	24 695	162.	I	60	vývoj	20 633
13.	I	5	čas	75 881	63.	F	33	noc	33 788	113.	F	51	stavba	24 665	163.	I	61	klub	20 459
14.	F	6	žena	75 329	64.	F	34	síla	33 742	114.	I	42	základ	24 607	164.	F	67	podpora	20 321
15.	M	2	muž	74 062	65.	I	20	vztah	33 700	115.	N	15	číslo	24 209	165.	I	62	tisíc	20 245
16.	I	6	případ	72 894	66.	F	35	noha	33 693	116.	I	43	smysl	24 205	166.	I	63	vzduch	20 225
17.	F	7	hlava	72 080	67.	I	21	program	33 630	117.	F	52	role	24 196	167.	N	22	zařízení	20 216
18.	N	2	oko	70 807	68.	I	22	hlas	33 265	118.	F	53	Evropa	24 104	168.	M	14	lékař	20 198
19.	N	3	město	70 241	69.	I	23	důvod	33 091	119.	F	54	skutečnost	24 086	169.	F	68	podoba	20 185
20.	F	8	cesta	68 611	70.	F	36	hra	33 083	120.	I	44	rozdíl	24 040	170.	I	64	úřad	20 180
21.	I	7	dům	64 372	71.	I	24	film	32 921	121.	F	55	zpráva	24 039	171.	F	69	součást	19 980
22.	F	9	věc	63 933	72.	F	37	minuta	32 882	122.	I	45	zdroj	23 959	172.	N	23	množství	19 920
23.	F	10	voda	62 399	73.	F	38	informace	32 455	123.	F	56	energie	23 828	173.	I	65	pohyb	19 895
24.	F	11	společnost	61 410	74.	I	25	prostor	32 176	124.	I	46	krok	23 791	174.	M	15	ředitel	19 871
25.	F	12	část	60 571	75.	F	39	pravda	32 032	125.	I	47	metr	23 697	175.	F	70	volba	19 757
26.	F	13	chvíle	58 435	76.	F	40	služba	31 914	126.	N	16	kolo	23 621	176.	I	66	názor	19 701
27.	I	8	systém	58 224	77.	I	26	druh	31 279	127.	N	17	%	23 511	177.	I	67	okamžik	19 596
28.	F	14	hodina	56 758	78.	N	8	století	31 278	128.	N	18	centrum	23 493	178.	F	71	myšlenka	19 556
29.	N	4	slovo	55 373	79.	F	41	Kč	30 673	129.	F	57	paní	23 441	179.	F	72	síť	19 422
30.	I	9	problém	54 415	80.	N	9	procento	30 315	130.	I	48	začátek	23 324	180.	I	68	plán	19 375
31.	F	15	cena	52 219	81.	F	42	ulice	29 884	131.	M/I	18	člen	23 277	181.	F	73	zkušenost	19 343
32.	F	16	Praha	52 029	82.	I	27	zákon	29 876	132.	F	58	akce	23 052	182.	N	24	rameno	19 310
33.	F	17	řada	50 694	83.	I	28	typ	29 847	133.	I	49	obraz	22 837	183.	F	74	úroveň	19 280
34.	F	18	škola	50 466	84.	F	43	hodnota	29 676	134.	F	59	forma	22 807	184.	I	69	jazyk	19 243
35.	I	10	konec	50 450	85.	I	29	pocit	29 591	135.	I	50	byt	22 788	185.	N	25	pole	19 225
36.	F	19	firma	49 113	86.	I	30	stav	29 496	136.	N	19	období	22 787	186.	F	75	miliarda	19 060
37.	F	20	koruna	45 414	87.	N	10	okno	29 070	137.	F	60	činnost	22 699	187.	I	70	les	18 985
38.	I	11	pohled	44 960	88.	N	11	právo	28 962	138.	M	7	Jan	22 607	188.	N	26	divadlo	18 943
39.	I	12	způsob	44 157	89.	I	31	počet	28 939	139.	M	8	přítel	22 553	189.	M	16	prezident	18 889
40.	F	21	dveře	43 294	90.	I	32	směr	28 924	140.	F	61	obec	22 537	190.	F	76	zahrada	18 871
41.	F	22	možnost	43 140	91.	F	44	smrt	28 328	141.	F	62	republika	22 478	191.	F	77	nemocnice	18 755
42.	F	23	oblast	42 030	92.	I	33	stůl	28 251	142.	I	51	proces	22 476	192.	I	71	vliv	18 740
43.	I	13	milión	41 189	93.	M	4	pán	28 114	143.	M	9	hráč	22 381	193.	F	78	policie	18 716
44.	I	14	peníze	40 618	94.	I	34	tým	27 971	144.	I	52	soud	22 114	194.	M	17	pes	18 698
45.	N	5	tělo	40 411	95.	F	45	vláda	27 851	145.	M	10	syn	21 939	196.	F	79	soutěž	18 689
46.	F	24	situace	40 242	96.	I	35	zájem	27 839	146.	I	53	materiál	21 924	195.	I	72	věk	18 689
47.	I	15	stát	40 005	97.	I	36	trh	27 515	147.	I	54	výkon	21 890	197.	N	27	zvíře	18 593
48.	F	25	otázka	39 932	98.	I	37	kraj	27 487	148.	I	55	rámec	21 881	198.	F	80	míra	18 587
49.	F	26	kniha	39 619	99.	I	38	bod	26 914	149.	I	56	vlas	21 618	199.	F	81	spousta	18 549
50.	F	27	skupina	39 589	100.	F	46	podmínka	26 663	150.	M	11	bůh	21 519	200.	I	73	prst	18 543

Tabulka 3.2 Frekvenční distribuce substantiv v SYN2015

Sloupec r uvádí rank, pořadí daného lemmatu všemi substantivními lemmaty, které SYN2015 obsahuje (tj. 159 804, na které připadá 29 549 285 jejich výskytů v různých tvarech), sloupec g udává rod (popř. rody) lemmatu (femininum (F), neutrum (N), maskulinum animatum (M) a maskulinum inanimatum (I)), sloupec p pořadí lemmatu podle rodu, sloupec lemma obsahuje reprezentativní tvar daného slova a sloupec f absolutní frekvenci lemmatu.

Z tohoto souboru byl následně vytvořen vzorek celkem 20 nejfrekventovanějších lemmat tak, (a) aby jich od každého rodu bylo zastoupeno pět, (b) aby byly zastoupeny i méně časté deklinační vzory, (c) aby vybraná lemmata nepředstavovala pro nerodilé mluvčí i následné vyhodnocování přílišné obtíže z důvodu nepravidelného skloňování či lexikální supletivnosti forem a (d) aby pak následně v korpusu CzeSL obsahovala alespoň minimální počet výskytů, (e) byla také vyloučena všechna propria.

Naplnění podmínky (a) vysvětluje, proč bylo třeba vybírat z takto rozsáhlého souboru lemmat – naprostou většinu mezi nejfrekventovanějšími lemmaty totiž tvoří substantiva rodu ženského a mužského neživotného. Ze sta nejfrekventovanějších je jich ženského rodu téměř polovina (46), mužského neživotného dalších 38, dohromady tedy obsahují 84 pozic. Životná maskulina jsou oproti tomu v nejčetnějším stu lemmat jen 4. I kdyby tedy všechna tato maskulina splňovala výše uvedená kritéria výběru, je třeba překročit hranici stovky.

Kritérium (b) je zvláště důležité, pokud bychom chtěli, byť na v tomhle ohledu velmi omezeném vzorku, pozorovat tendence chybování v „nedominantních“ vzorech přebíráním koncovek vzorů dominantních. Schöne (2015) dokládá určující postavení vzoru žena (zvláště pro feminina, ale nejen pro ně) a vzoru hrad (pro neživotná i životná maskulina). V případě neživotných maskulin tak například bylo vybráno slovo *konec* – první lemma tohoto rodu (celkově 35. a mezi neživotnými maskuliny 10.), které splňovalo všechna ostatní kritéria a skloňuje se podle vzoru *stroj*.

Na základě podmínky (c) bylo například vyloučeno lemma *dítě*, jehož singulárové formy náleží k jinému rodu než formy plurálové, lemma *člen*, které má životnou i neživotnou variantu, lemma *ruka*, *člověk*, *rok* atd.

Kritérium (d) vedlo například k upřednostnění lemmatu *lékař* před lemmatem *hráč*. Ačkoli *hráč* je v SYN2015 celkově 143. lemma (a 9. mezi životnými maskuliny), v korpusu CzeSL-SGT je

zastoupen jen 17 výskyty. Byl proto raději vybrán *lékař* 168. celkově (14. mezi M), který má v korpusu nerodilých mluvčích 78 výskytů. Kvůli frekvenci byla lemmata vylučována pouze výjimečně, v případech, kdy bylo zřejmé, že z počtu jejich výskytů v CzeSL by se stěží dalo něco vyvozovat. Nejnižší frekvenci ze slov, která byla zahrnuta do vzorku, má právě slovo *lékař*.

Bod (e) se týká lemmat *Praha, Evropa, Jan, Petr*, pravděpodobně též *republika*. Propria mají ze své podstaty odlišnou distribuci tvarů, například se zpravidla neobjevují v plurálových formách, proto nebyla do analýzy zahrnuta.

Výsledný vzorek obsahoval následující lemmata:

č	r	p	f	lemma
1	4.	1	108 001	doba
2	6.	3	96 721	život
3	7.	1	95 819	místo
4	9.	3	81 296	práce
5	10.	4	79 949	svět
6	11.	4	79 801	strana
7	12.	5	77 054	země
8	13.	5	75 881	čas
9	15.	2	74 062	muž
10	16.	6	72 894	případ
11	19.	3	70 241	město
12	24.	11	61 410	společnost
13	29.	4	55 373	slovo
14	35.	10	50 450	konec
15	51.	3	38 344	otec
16	57.	6	36 437	jméno
17	78.	8	31 278	století
18	145.	10	21 939	syn
19	153.	12	21 065	autor
20	168.	14	20 198	lékař

Tabulka 3.3 Frekvence lemmat v SYN2015

Sloupec č značí číslo lemmatu ve vzorku, r opět rank v SYN2015, p pořadí podle rodu, f absolutní frekvenci v SYN2015, rod lemmatu je znázorněn barvou.

3.2.1 Vyhledávání v CzeSL

K těmto vybraným lexémům byly následně z korpusu CzeSL-SGT získány všechny výskyty příslušného lemmatu. Vzhledem k nepřesnosti automatické lemmatizace a možnosti vyhledávat na dvou rovinách anotace (viz oddíl 3.1.2) byl k tomuto účelu zvolen následující dotaz: [lemma="příklad" | lemma1="příklad"], cílem je tedy vyhledat všechny výskyty určitého slova (word), které byly před automatickou emendací přiřazeny k danému lemmatu (lemma), nebo všechny výskyty určitého slova (word1), které byly k danému lemmatu přiřazeny až po automatické emendaci daného tvaru (lemma1). V případě, kdy dané lemma může mít v češtině ještě jinou slovnědruhovou platnost než substantivní, byl dotaz doprovozen specifikací pomocí tagu, např.: [lemma="místo" | lemma1="místo"].

Motivací k formulaci takového dotazu byl korpusový průzkum případů, které a) zachytí automatická lemmatizace opraveného tvaru (word1), ale nezachytí automatická lemmatizace původního tvaru (word1), b) zachytí automatická lemmatizace původního tvaru (word), ale nezachytí automatická lemmatizace opraveného tvaru (word1),

Ilustrováno na příkladu lemmatu *doba* by seznam tvarů odpovídající situaci a) a získaný dotazem [lemma1="doba" & lemma!="doba"] vypadal následovně:

č.	word	f
1	tobou	14
2	dobe	8
3	dobé	7
4	dobý	2
5	dobá	2
6	dobr	2
7	tudobu	1
8	tobóu	1
9	sobou	1
10	hoby	1
11	dóba	1
12	dubou	1
13	dově	1
14	dola	1
15	dobých	1
16	dobyi	1
17	dobi	1

18	dobach	1
19	do13	1
20	debě	1
21	Tobou	1

Tabulka 3.4 Frekvenční distribuce tvarů lemmatu *doba* (podle atributu word)

Z celkem 50 získaných tvarů se ve 22 případech jedná o jiné slovo než *doba*, ve 2 případech to nelze rozhodnout a ve 26 případech opravdu jde o tvar slova *doba*, přibližně v polovině případů tedy automatická lemmatizace původního tvaru patřičné lemma nezachytila. Jde o případy s chybnou diakritikou i více vzdálené tvary (např.: *V součastně dově čas je jedna z nějvětších hodnot*). Při formulaci dotazu je tedy nutné zahrnout lemma1, aby z analýzy nebyly vyloučeny všechny tvary s diakritickou nebo pravopisnou chybou nebo morfologicky chybné tvary, pokud mají podobu tvaru neexistujícího.

Pokud jde o opačnou situaci b), výsledkem dotazu [lemma="doba" & lemma1!="doba"] je prázdná množina. Neexistují tedy tvary, které by v původním textu byly lemmatizovány jako *doba* a po opravě tvaru by tak lemmatizovány nebyly. To ovšem neplatí vždy. Stejný dotaz pro lemma *město* vygeneruje následující výskyty:

(1) ROMÁNU ANDRIĆ CHRONOLOGICKY POPISUJE KAŽDODENNÍ ŽIVOT < MĚSTA /město/MOST/most/NNIS1-----A----/S/SingCh > VIŠEGRADU , KTERÝ SE NACHÁZÍ NA BŘEHU ŘEKY DRINY

(2) každý pracovný den . poslouchan radio v 6 hodin , < m /město/m/metr/NNIXX-----A---8// > . V uterý veceř na televizi je moc zajimovy , česky

(3) Vám ještě pozdě . S pozdravem . Váš syn Adam MOJE < MĚSTO /město/MOST/Most/NNIS1-----A----/S/SingCh > . Moje město jmenuje se Moskva . To je velké , krasné

(4) KOUPELNA A ZACHOD . POLOHA BYTU : CENTRUM HLAVNIHO < MĚSTA /město/MSA/MSA/NNNXX-----A---8/S/SingCh > . PŘIBLIŽNĚ CENA : 10000000 Kč . TELEFON : 123456

V příkladu (1) je lemma město nesprávně opraveno na lemma *most*, stejně tak v příkladu (3), v příkladu (4) je nesprávně opraveno na neexistující lemma *msa* a v příkladu (2) lemma nelze určit, pravděpodobně se ale nebude jednat ani o *město*, ani o *metr*.

V případě slova *den* jsou u správných tvarů jako lemma1 uváděna slova *dno* a *dna*, u správného tvaru *lidmi*, místo lemma *člověk* lemma *lima* atd. V automatické lemmatizaci

opraveného tvaru (resp. opravě původního tvaru) se tedy vyskytují chyby, které v původní verzi nejsou.

Z výše uvedených a) a b) příkladů plyne, že pro pokud možno co nejpřesnější frekvenční obrázek o určitém slovu a jeho chybovosti je optimální použití dotazu [lemma="příklad" | lemma1="příklad"].

V případech, kdy je v korpusovém rozhraní, např. pro výpočet frekvenčních distribucí, nutné volit mezi emendovanou a neemendovanou rovinou (to se týká atributů word, lemma i tag), volím rovinu emendovanou. Gramatické profily tedy např. budu porovnávat na základě opravených tagů (viz dále) a absolutní frekvence lemmat vybraných pro analýzu v CzeSLu je v následují tabulce o něco nižší, než bude počet výskytů zahrnutých do manuální anotace po připočtení *lemma* k *lemma1*:

č	r	p	f SYN	lemma	f CzeSL
1	4.	1	108 001	doba	740
2	6.	3	96 721	život	3010
3	7.	1	95 819	místo	1081
4	9.	3	81 296	práce	1109
5	10.	4	79 949	svět	891
6	11.	4	79 801	strana	258
7	12.	5	77 054	země	1059
8	13.	5	75 881	čas	1598
9	15.	2	74 062	muž	1191
10	16.	6	72 894	případ	123
11	19.	3	70 241	město	3344
12	24.	11	61 410	společnost	197
13	29.	4	55 373	slovo	415
14	35.	10	50 450	konec	276
15	51.	3	38 344	otec	340
16	57.	6	36 437	jméno	162
17	78.	8	31 278	století	145
18	145.	10	21 939	syn	302
19	153.	12	21 065	autor	86
20	168.	14	20 198	lékař	78

Tabulka 3.5 Frekvence lemmat v CzeSL-SGT (podle atributu lemma1)

3.3 Anotace

Korpus CzeSL-SGT je anotován automaticky (viz výše). V praxi to znamená, že spolehlivě rozpozná takové chyby, kdy je užit takový tvar lemmatu, který je nějakým způsobem (ve většině případů diakriticky, nikoli však nutně morfologicky) defektní. Chyby, které spočívají v kontextově chybném užití existujícího tvaru, rozhodně nejsou rozpoznávány nijak spolehlivě.

Například v případě lemmatu *konec* to znamená, že z celkem 21 chyb spočívajících v užití nesprávného, ale v paradigmatu daného substantiva existujícího tvaru a neobsahujících diakritickou chybu, jich je automatickou anotací jako chybných označeno 5, a to v následujících případech:

(5) a Rusko mají kousek společné historie . Tak už na < konce > mého stáže jsem věděla , že chci žít a pracovat v

(6) se snaží dostat na VŠ . Do Prahy jsem přijela už na < konce > září a hned jsem přistoupila k vyučování . Na začátku

(7) v Praze . Učím se češtinu v Praze už 8 měsiců a na < konce > května bude moje studium ukončeno . Když jsem čekala

(8) žít . Jsem studentka , a teď bydlím v Praze . Ale do < konci > jsem ješte nerozhodla , kde chtěla bych dokončit

(9) Praha mi přinesla " pražské archivy . " Ke < konců > roka musim napsat magisterskou prace o československo

Všechny tyto případy jsou ohodnoceny značkou „G/SingCh“, což znamená, že tyto tvary byly vyhodnoceny jako gramaticky chybné (G), a to z důvodů záměny jednoho znaku za druhý.[27]

Zároveň korpus ale obsahuje řadu obdobných případů, které jako chybné hodnoceny/označeny nejsou, například:

(10) fotografovali hodně . Seznamila jsem se s novými lídmi . Na < konce > dovoleny jsem koupila různé dárky u . Byla to moc

[27] http://utkl.ff.cuni.cz/~rosen/public/SeznamAutoChybR0R1.html

(11) tam taky znám dobrou restaurace , budeme večeřet Na < konce > dnu ve 20 hodin pujdeme na jazz koncert do klubu

(12) mě a ještě někdy psalá na tabule duležite věci . Na < konce > přednáški začala přednášet doktorantka . To byla

(13) jsem býl malý a to bylo v dětství . Vždy jsem byl na < konce > třidy , ale to nebylo tak strašně , protože tři z

Příklady (5)–(7) představují typově stejné chyby jako příklady (10)–(13), tvar *konce* je v nich užit namísto lokálu singuláru. Důvod, proč automatická anotace první sadu příkladů jako chybnou označila a druhou nikoli, se odvíjí od odlišnosti tagů. Tvary v (5)–(7) byly automaticky označeny jako lokál singuláru, (10)–(13) jako akuzativ plurálu a jako akuzativ plurálu je tvar *konce* správný.

Automatické značkování pádu a čísla představuje dalších problematický okruh. Chybovost v jejich vyhodnocení zdaleka přesahuje 4% hladinu chybovosti, která se uvádí pro korpusy SYN (srov. Schöne 2015). Zůstaneme-li u příkladu lemmatu *konec*, je z celkem 276 výskytu 80 tagováno špatně.

V případě lemmatu *konec* tedy úspěšnost identifikace chyby dosahuje sotva 25 % a pravděpodobnost toho, že tag opraveného tvaru (tag1) bude označovat takový pád a číslo, které by v daném kontextu byly správné, 30 %. Protože spolehlivost anotace chyby i tagu jsou pro účely této práce směrodatné, vyhodnotila jsem jako nezbytné projít všechny výskyty vybraných 20 lemmat ručně.

3.3.1 Cíle, zásady, problémy

Cílem ruční anotace bylo identifikovat morfologické chyby v užití vybraných substantivních lemmat. Vycházela jsem přitom z minimální cílové hypotézy, tak jak je formulována pro účely korpusu Merlin (viz výše TH1, srov. Schöne 2015): cílem je gramaticky správná, nikoli nutně standardní věta, značí se chyby ortografické a morfologické. Zároveň jsem se zaměřila pouze na KWIC. Mimo zřetel tedy stály jak vyšší jazykové roviny, tak ostatní jazykové jednotky.

Jako správné jsou tedy hodnoceny i následující, sémanticky poněkud neobvyklé případy, ve kterých měl mluvčí na mysli pravděpodobně jiné slovo:

(14) Bratr je student , studuje na střední škole . Můj < otec > se jmenuj Eva , pracuje v nemocnici . Sportoval XXX

(15) on končil univerzitu . Musel udělat hodně úkolů pro < konec > univerzity . Každý den chodil a studoval matematiku

Stejně tak stojí mimo zájem analýzy případy, kdy se morfologická chyba nachází v syntagmatu, ale nikoli přímo v KWIC, například chyby ve slovesné třídě (16), (17) nebo ve skloňování adjektiv (18):

(16) Maminka diví na televize , a oteč ctě knihy . Jejich < syn > spije . Vlevo na stene je velky a krasny obráz .

(17) Samozřejmne , popsaná situace je velmi smutná , ale < autor > vyvoláje naději , že nová generace dosa sklidit úspěch

(18) je Dmitrií Kantemir - velmi chytrý a intilegentný < muž > ve své době . On byl synem krála Moldavska - Antioha

Výjimku z výše uvedeného omezení se na ortografickou a morfologickou správnost KWIC ovšem tvoří případy, kdy význam věty nebo význam věty a chyba v kongruenci KWIC např. s adjektivem nebo se slovesem jasně ukazují, že užitý tvar paradigmatu je v daném kontextu nesprávný.

Například u následujícího dokladu (19) ze sémantických důvodů nedává plurál lokálu smysl (*rodiným městem* je pravděpodobně míněno rodné město a i v případě, že by se mělo jednat o město, kde se nachází rodina mluvčího, bylo by pravděpodobně pouze jedno) a pro singulárovou formu mluví i přívlastky v podobě přivlastňovacího zájmena a adjektiva:

(19) českého jazyka je lehky pro Japonci . v mém rodiném < městech > je české restaurace a tam pracujou češky . Navštívila

Tento příklad byl tedy označen jako morfologická chyba. Pro srovnání uvádím následující příklady, ve kterých by se singulárová forma mohla též jevit jako vhodnější, resp. by se ze širšího kontextu dalo domýšlet, že intence mluvčího byla pravděpodobně referovat pouze k jednomu *místu* a jedné *společnosti*, neobsahuje pro to ale dostatek indicií, a příklady (20) a (21)(16) jsou tak hodnoceny jako správné.

(20) plus v tom , že můžu mluvit s Vadimem o Petrohradú . < Místech > kde nikdo kromě nás nikdy nebyl . Ale to neruší mych

(21) nejdůležitejší , protože má máma pracuje nyní jako ředitel < společností > ochrany a pomoci zviře . Ve měste Kazan má rodina

Pokud byla v rámci syntagmatu narušena shoda, ale není možné jednoznačně určitě, že ve špatném tvaru stojí KWIC, nikoli druhý člen syntagmatu, jako určující je chápáno řídící slovo, tedy sledované substantivum, a řádek není hodnocen jako chybný, případy (22), (23) a (24) tak například nejsou považované za morfologickou chybu v rodě, jakkoli se u první dvou může jednat o chyby pod vlivem vzoru *žena* a posledního podle měkkých mužských vzorů, případ (25) není považován za chybu v čísle a (26) za chybu v pádu:

(22) proto mám rád čestovat a chtěl bych čestovat po celé < světě > . Byl jsem v střední škole , nemá peníze . Takže

(23) Řídí náš život reklama ? Reklama je všude na naši < světě /svět/světě/svět/NNIS6-----A----// > . Nejenom vidíme z televize , z Internetu , ale taký

(24) kteří tady žijou . Rozumět lidi je moc důležité v mém < práci /práce/práci/práce/NNFS6-----A----// > jako kněz . Co se nejvíc změní ? Asi nic nebo nic

(25) modernější , ale nevím jak pracuje tady politické < strany > . Potom , v Turecku nemužete vidět mini sukně , protože

(26) důležitější než obory humanitní . Protože v současnou < době /doba/době/doba/NNFS3-----A----// > máme velké problémy s ekologií , se spotřebou energii

Specifický problém představují substantiva *čas* a *práce*, která se v určitých svých významech chovají jako nepočitatelná. Příklady jako (27) a (28), kdy je po nějakém kvantifikátoru užita forma genitivu plurálu, jako chybné značeny nebyly, zároveň k těmto příkladům ale bude přihlédnuto při analýze vlivu gramatických profilů na chybovost ve tvarech těchto lemmat.

(27) těžké . Mám jenom jeden problem : potřebuju hodně < časů > , když chci dělat palačinku . Mám ráda taky palačinku

(28) Pišu tě dopis pro můj život v Praze . Teď mám hodně < prací > - studuju češtinu na univerzite . Každý dén chodím

Při rozhodování mezi označením chyby jako diakritické/pravopisné nebo jako morfologické v případech, kdy by odstraněním/doplněním diakritických znamének vznikl gramaticky správný tvar, bylo zvoleno řešení, které se odchyluje od výše uvedeného algoritmu (viz oddíl 3.1.1), který byl na základě českých příkladů vytvořen pro anotační schéma projektu Merlin. Ačkoli už algoritmus Štindlové a kol. (2014) je zmírněním původní zásady anotace Merlin značit všechny chyby v koncovce flektivních jmen jako gramatické – podle něj jsou jako gramatické vyhodnocovány jen případy, kdy lze daný tvar považovat jak za chybu ortografickou, tak za gramatickou, tj. daný tvar existuje v paradigmatu slova –, domnívám se, že i to vede k poměrně velkému zkreslení. V následujících případech by například tvar (29) byl hodnocen jako gramaticky chybný, zatímco tvar (30) nikoli, přestože se jedná o podobné příklady; o tom, že se v příkladu (29) nejedná o záměnu tvar singuláru dativu s instrumentálem nebo plurálem genitivu, ale o chybu v kvantitě vokálu, svědčí navíc i okolí KWIC – stejnou chybu obsahuje i *práci* předcházející předložka *kvulí* (kvůli).

(29) špatné veci . Myslím , přiroda bude špinavejí kvulí < prací > člověka a rozvoje světu . Ale kdykolív budeme starát

(30) strašná , a chtěla zabít každého , koho viděla . Na < koncí /koncí/konci/konec/NNIS6-----A----/S/Quant1 > filmu hlavní herec zabil mumiju a oženil se na krasně

S ohledem na četnost diakritických chyb celkem i jejich četnost v případech, kdy gramatická chyba není tou pravděpodobnější variantou, a chyb ve srovnatelných případech, kde ale dané tvary v paradigmatu slova neexistují, byla pro účely této práce zvolena ještě „mírnější“ varianta anotace.

V případech, kdy je pravděpodobnější, že se jedná o chybu ortografickou, jako např. ve (29), hodnotím chybu jako ortografickou, a naopak, v případech, kdy chybí indicie pro takové rozhodnutí, byly tvary ohodnoceny oběma značkami (jako gramatické i jako ortografické) a v analytickém oddíle jsou mezi morfologické chyby počítány jen tam, kde je to explicitně uvedeno.[28]

Kromě chyb, které byly podle výše uvedených zásad hodnoceny jako diakritické, pravopisné (všechny ostatní mimo diakritických, nejčastěji y/i) a morfologické, byly všechny (správné i nesprávné) tvary ohodnoceny tagem2, tedy v zásadě opravou opraveného tagu (tag1), který značil pád a číslo tvaru (případně též variantní koncovky) po vzoru tagů ÚČNK. V dalších krocích byly řádky označené jako chybné vyselektovány k podrobnější anotaci (3.3.3) a tagy2 a počet ortografických a morfologických chyb byly usouvztažněny s gramatickými profily (3.3.2).

3.3.2 Sestavování profilů

Gramatické profily byly pro účely této práce převzaty z výzkumu Lehečkové, Lázničky a Jandy (2016), jedná se o pořadí frekvenční distribuce pádů pro určitá lemmata, získaná z korpusu SYN2015. Pro každé lemma tohoto vzorku byly tyto profily usouvztažněny s informacemi získanými z korpusu CzeSL a z ruční anotace vzorku z tohoto korpusu, jak shrnuje například následující tabulka pro lemma *doba*.

[28] Jsem si zcela vědoma značné subjektivnosti takové klasifikace, domnívám se ale, že výsledné zkreslení bude nižší, než kdyby všechny případy, kdy se může jednat o diakritickou/pravopisnou chybu, byly hodnoceny jako gramatické jen proto, že užitý tvar se nachází v paradigmatu slova.

A	B	C	D	E	F	G	H	I	J	K
lemma	tag	tvar	f SYN	SYN	f CzeSL 1	CzeSL 1	f CzeSL 2	CzeSL 2	err	%
doba	nnfs6-----a-----	době	40339	1	342	1	367	1	23	6,3
doba	nnfs4-----a-----	dobu	24905	2	139	2	135	2	6	4,4
doba	nnfs2-----a-----	doby	22811	3	101	3	96	3	2	2,1
doba	nnfs1-----a-----	doba	7920	4	82	4	81	4	4	4,9
doba	nnfs7-----a-----	dobou	4334	5	25	5	7	6	0	0,0
doba	nnfp2-----a-----	dob	3366	6	12	7	10	5	0	0,0
doba	nnfp6-----a-----	dobách	2503	7	6	8	6	7	2	33,3
doba	nnfs3-----a-----	době	773	8	22	6	1	10	0	0,0
doba	nnfp1-----a-----	doby	592	9	6	9	2	9	0	0,0
doba	nnfp4-----a-----	doby	355	10	5	10	3	8	1	33,3
doba	nnfp3-----a-----	dobách	50	11			0		0	
doba	nnfp7-----a-----	dobami	49	12			0		0	
doba	nnfs5-----a-----	dobo	3	13			0		0	
doba	nnfp5-----a-----	doby	1	14			0		0	

Tabulka 3.6 Frekvenční profil lemmatu *doba*

Sloupce A–C, obsahující atributy lemma, tag a word/tvar, jsou shodné pro oba korpusy,[29] další dva sloupce uvádějí frekvenci daného tvaru (D) a jeho pořadí (E), následující čtyři sloupce obsahují tytéž informace pro korpus CzeSL-SGT, v případě F a G tak, jak je zachycuje automatická anotace korpusu, sloupce H a I jsou výsledky ruční anotace vzorku, sloupec J udává kolik chyb bylo ve vzorku pro daný tvar identifikováno a sloupec K počet těchto chyb vyjadřuje procentuálně (vzhledem ke sloupci H).

3.3.3 Chybová anotace

Podrobnější klasifikace chyb byla převzata z pilotního výzkumu Karin Schöne (2015), jež sama zmiňuje potřebu ověřit výsledky své analýzy na rozsáhlejším korpusovém vzorku. Od zásad jí navržené anotace se tedy snažím pokud možno nelišit.[30]

Schöne anotaci provádí na čtyřech rovinách. Rovina 1 vymezuje pozici, na které k chybě došlo, a to jako bezprostřední složku (s typy a podtypy NP, NP-kvant, NP-adv, VP, PP-adv a AdP-kvant). Rovina 2 stanovuje druh chyby se čtyřmi možnostmi: chyba v přeložce, neexistující tvar,

[29] Seznam tagů byl v korpusu CzeSL kvůli přiřaditelnosti k tagu ze SYN2015 získán na základě atributu tag1.

[30] Mnohdy to všem komplikuje skutečnost, že některé klasifikační „zádrhele“ nejsou dodatečně popsány.

morfologicky nesprávný tvar (tj. tvar jiného pádu téhož paradigmatu) a pravopisná chyba. Rovina 3 pro příklad morfologicky nesprávného či chybného tvaru specifikuje, jakou morfologickou kategorii chyba zasahuje (pád, číslo, rod), a rovina 4, explanační, se snaží určit, na jakém základě k chybě došlo (podle jiného pádu nebo vzoru nebo užitím variantní koncovky, která se pro daný lexém použít nedá).

Schematicky shrnuje Schöne (2015: 25) poslední tři roviny takto:

Rovina 2	Rovina 3	Rovina 4
Chyba v předložce	nesprávná volba předložky (wrong choice) nadbytečná předložka (addition) chybějící předložka (omission)	-
Neexistující tvar	chybný rod nelze určit	jiný vzor jiný vzor anebo jiný pád
Morfologicky chybný tvar	chybný pád chybné číslo chybný rod nelze určit (ambiguous) nesprávná varianta koncovky	jiný vzor jiný pád jiný vzor anebo jiný pád jiná varianta koncovky
Pravopisná chyba	nesprávná volba grafému chybějící grafém (omission) nadbytečný grafém (addition)	-

Tabulka 3.7 Přehled rovin anotace (Schöne 2015)

Na základě tohoto modelu byly ručně anotovány všechny případy morfologicky chybného tvaru, neexistujícího tvaru a chyby v předložce,[31] tj. celkem 790 korpusových dokladů 20 zkoumaných lemmat. Čistě pravopisné chyby dále anotovány nebyly, jednak proto, že dohromady čítají dalších přibližně 1 500 tokenů, jednak proto, že analýza toho, zda se v případě pravopisné chyby jedná o nesprávnou volbu grafému, chybějící grafém či grafém nadbytečný nestojí v centru zájmu této práce, v naprosté většině případů nadto pravopisná chyby znamená chybu diakritickou, a tím pádem i značku nesprávná volba grafému.

Při anotaci vybraných chyb se jako zvláště problematické ukázalo rozhodování o tom, jaký typ chyby přiřadit morfologicky chybnému tvaru, tedy zda jej hodnotit jako chybu v pádu, číslu nebo rodu substantiva. Aby se do anotování nepromítalo příliš vlastní interpretace a očekávání, jaká jiná

[31] Při posuzování správnosti/existence tvaru bylo za směrodatné považováno paradigma slova tak, jak jej uvádí Internetová jazyková příručka, dostupná z: http://prirucka.ujc.cas.cz/.

forma (pád, číslo, rod) bude hrát v chybném tvaru roli, byly danému tvaru přiřazeny všechny možnosti.

Případy (31)–(33) tak například byly označeny jako chyba v pádu i chyba v rodu, příp. vzoru. V příkladu (31) mohla být příčinou chyby volba nominativu či genitivu singuláru namísto lokálu i přiřazení země ke vzoru žena, stejně tak v případě (32) mohl být genitiv singuláru zaměněn za dativ, nebo (pod)vzor les za vzor hrad, v příkladu (33) mohla být chyba motivována dativem či lokálem singuláru nebo přiřazením slova *muž* k měkkému ženskému vzoru růže.

(31) Moje nejhezčí dovolená , když jsem byla poprve v cizí < země > . To bylo Turecko . V tu dobu mi bylo sedm let .
(32) problemy a stale s nimi bojuje , je to cyklus našeho < zivotu > . Tak musíme se snažít uživat ho jak je možne vic
(33) pověsí venku na dveře jako ochrana domu a rodiny . Pro < muži > hezký suvenýr mohl by malý kámen z okraje Sahary

Zároveň ale bylo více než při selekci chybných výskytů z původního vzorku (viz 3.3.1) přihlíženo k okolí KWIC. To znamená, že pokud se nějaká z možností jevila jakkoli pravděpodobnější, byla označena pouze ona.

Morfologicky chybný tvar v příkladu (34) je tak na rovině 3 značen pouze jako chyba v rodu a na rovině 4 je specifikován vzor moře, toto označení se jeví jako pravděpodobnější než chyba v pádu (zvolení tvaru pro nominativ či genitiv singuláru místo akuzativu) na základě ukazovacího zájmena ve středním rodě, v příkladu (35) je tvar opět hodnocen pouze jako chyba v rodu a skloňování *světa* podle vzoru žena kvůli adjektivními přívlastku, který je kongruentní s ženským rodem, a příklad (36) je zase označen pouze jako chyba pád (nominativ či instrumentál plurálu místo genitivu plurálu), nikoli jako chyba v rodu (vzor kost), protože z parataktického spojení je patrné, že daný tvar by měl stát v plurálu (*u muži* a *u žen*).

(34) toho mě musilo by stačit , abych rozumět , mám rád to < země > nebo ne rád . teďko mužů určitě řeknut , že českou

(35) protože si myslím , že prážský hrad je nejhezčí z celé < světy > , a proto ho musíme uvidět spolu , přestože už jsi

(36) velmi špatné ekologie . Proto průměrná delka zivota u < muzi > asi 61 let , a u žen asi 68 let . Take v Rusku máme

Příklad (36) zároveň naráží na další princip kterou je zásada jednoduché (jednokrokové) cesty k cílovému tvaru. To znamená, že jsou vždy uvedeny pouze varianty, ke kterým lze dojít buď změnou pád, nebo rodu, nebo čísla. Varianta, která by obnášela tvar interpretovat například jako chybu v pádu a zároveň v čísle je uvedena pouze v případě, že jednodušší výklad není možný. Bez této zásady by vzhledem k rozsáhlé homonymii koncovek vzniklo těžko analyzovatelné množství chybových hypotéz pro většinu tvarů.

4 Analýza

4.1 Chybovost v korpusu CzeSL

Tento oddíl kvantifikuje základní charakteristiky 20 vybraných a ručně anotovaných lemmat z korpusu CzeSL a srovnává je s výsledky pilotního výzkumu Karin Schöne (2015), ze kterého proto přebírá sledované kategorie.

4.1.1 Rovina 2 – typ chyby

Z celkem 16 504 výskytů sledovaných lemmat jich bylo 2 313 vyhodnoceno jako chybných. Tabulka 4.1 shrnuje rozložení chyb podle jejich typu. Více než polovina chyb (63,4 %) byla pravopisného charakteru, v necelých pěti procentech (4,6 %) se jednalo o chybu v předložce. Morfologicky (resp. morfosyntakticky) nesprávných tvarů vzorek obsahoval celkem 32,1 % – z toho se v 29,2 % jednalo o morfologicky chybný (ale v paradigmatu existující) tvar, v 2,9 % o tvar neexistující. Tyto dva typy chyb byly dále analyzovány podrobněji.

typ chyby	absolutní počet chyb	%
morfologicky chybný tvar	675	29,2 %
chyba v předložce	107	4.6 %
neexistující tvar	66	2,9 %
pravopisná chyba	1466	63,4 %
CELKEM	2314	100 %

Tabulka 4.1 Rozložení chyb podle typu

4.1.2 Rovina 1 – bezprostřední složky

Rozložení chyb podle bezprostředních složek ukázalo, že nejvíc chyb (41,8 %) připadá na předložkové fráze, přičemž nebylo rozlišováno, zda se jedná o adverbiální obligatorní či fakultativní doplnění (viz příklady (37) a (38)). Ve složce verbální, tedy v substantivech v pozici přímého či nepřímého objektu se chyba vyskytla ve 29,2 % případů ((39), (40)). V nominálních frázích se objevilo 20,1 % chyb, nejčastěji v pozici neshodného přívlastku ((41)). V adverbiálních a nominálních složkách obsahující kvantifikující výraz, které byly sledované zvlášť a vyčíslené dohromady, k chybě došlo v 8,9 % případů ((42), (43)).

(37) , protože eto dobře pro zdraví . Pak ja pojedu do < prací > , tám ja pracuju na počitače nebo pišu dokumenty

(38) je daleko . To bude zima . Budu jít na procházku do < město > , bruslit a lyžovat . Budu hrat na kytare . Vezmu

(39) potom jsme chtěli pryč . Když jsme zkusili obejít toho < muži > , on zavolal jeho kolegy a nás zablokovali . Vlastně

(40) říkala o životě Kanady , proto jsem se zajímala o < životě > cizí země . Kamarádka na mě způsobila . Teď bydlím

(41) město v Evropě . Rada se procházím ulicemi tohoto < město > v Evropě , protože v každé ulice mohu narazit na

(42) jsem studovala česky na univerzitě ne mám moc volne < čas > . Každy dopoledne do 9:00 na 13:00 jsem v škole ale

(43) to bude obtízně . Peníze jsou velká čast lidského < životu > . Když budeme mluvit o budoucnosti , přemyslíme o

fráze	absolutní počet chyb	%
NP	149	20,1
PP	310	41,8
VP	216	29,2
NP/AdP-kvant	66	8,9
CELKEM	741	100 %

Tabulka 4.2 Rozložení chyb podle fráze

4.1.3 Rovina 3 – bližší určení druhu chyby

Rovina 3 ukazuje, v čem konkrétně a v čem nejčastěji morfologická chyba (kontextově chybný nebo neexistující tvar) spočívala. Tabulka 4.3 Rozložení „jednoznačných“ druhů chybtyto informace uvádí pro případy, u nichž nejjednodušší (jednokrokový, viz oddíl 3.3.3) postup nabízel jen jednu intepretaci chyby, tedy tvar hodnotil buď jako chybu v pádu, rodu, vzoru, nebo čísle. Tabulka 4.4 Rozložení „víceznačných“ druhů chybzobrazuje totéž pro případy „víceznačné“, tj. ty, v nichž se nebylo možné jednoznačně rozhodnout, jaké morfologické kategorii chybu přiřadit ((31), (32), (33)).

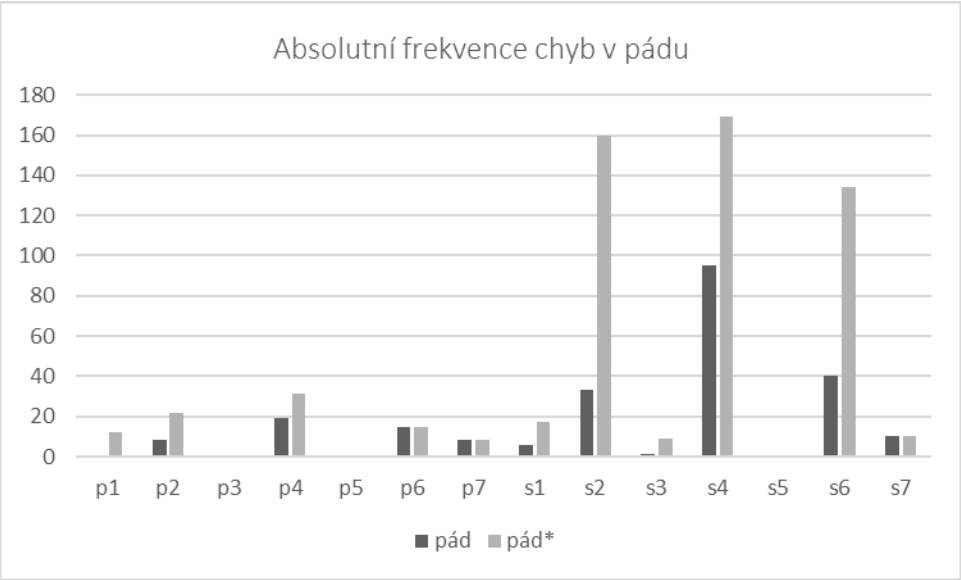
druh chyby	absolutní počet chyb	%
chybný pád	236	68,0
chybný rod	21	6,1
chybný vzor	48	13,8
chybné číslo	32	9,2
nelze určit	10	2,9
CELKEM	347	100 %

Tabulka 4.3 Rozložení „jednoznačných“ druhů chyb

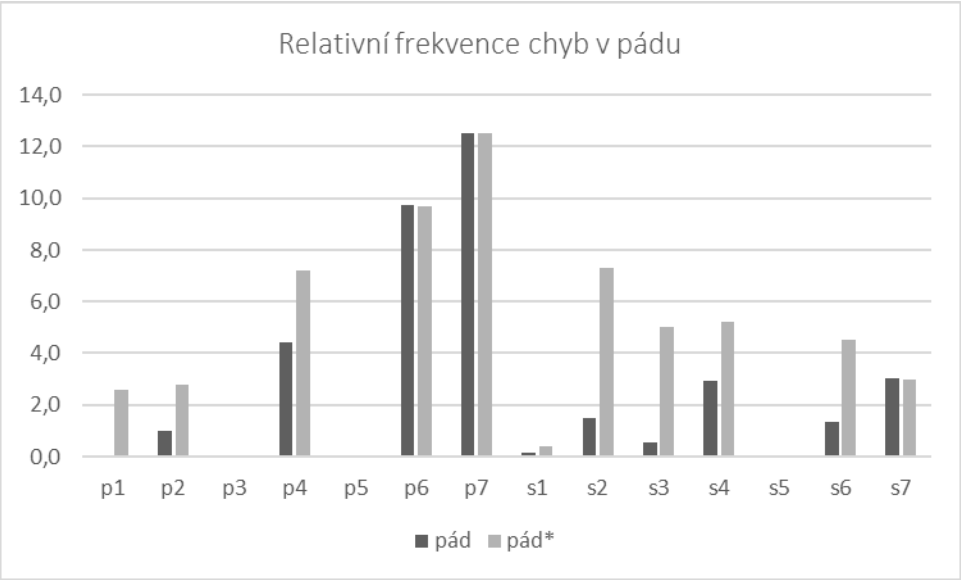
druh chyby	absolutní počet chyb	%
pád/rod	170	43,1
pád/vzor	116	29,4
pád/číslo	58	14,7
vzor/rod	16	4,1
vzor/číslo	0	0,0
číslo/rod	0	0,0
více možností	34	8,6
CELKEM	394	100 %

Tabulka 4.4 Rozložení „víceznačných“ druhů chyb

Většina chyb se týká morfologické kategorie pádu (v jednoznačných případech v 68 %, ve víceznačných lze jako chybu v pádovém tvaru interpretovat až v 87 % případů). Tabulka 4.5 shrnuje rozložení těchto chyb podle v jednotlivých pádových tvarech v absolutních číslech, Tabulka 4.6 v relativních (vzhledem k celkové frekvenci pádového tvaru ve vzorku). První, černé sloupce toto zachycují pro jednoznačné případy, druhé, šedé pro případy, kdy chyba nabízí více interpretací.

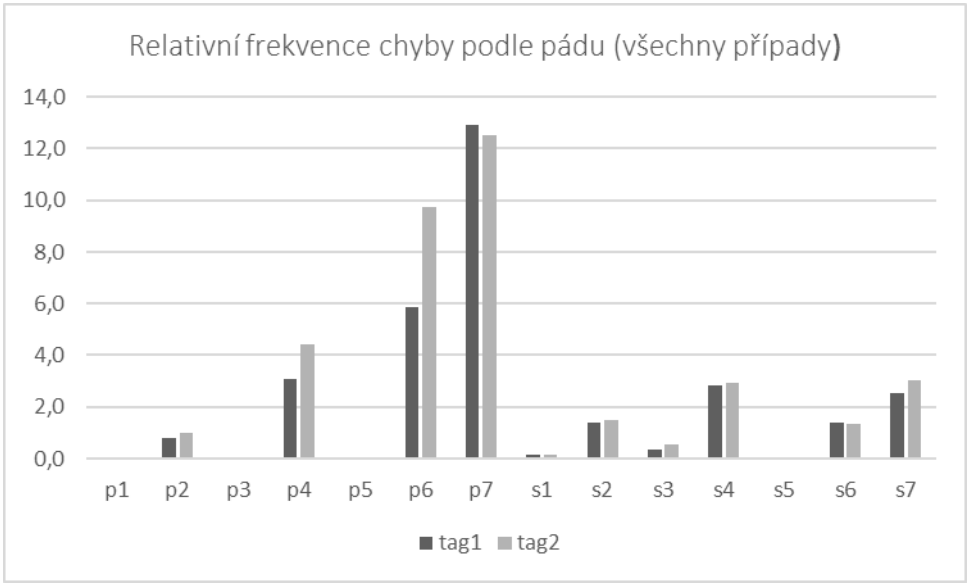

Absolutní frekvence chyb v pádu
180
160
140
120
100
80
60
40
20
0
p1
p2
p3
p4
p5
p6
p7
s1
s2
s3
s4
s5
s6
s7
pád
pád*

Tabulka 4.5



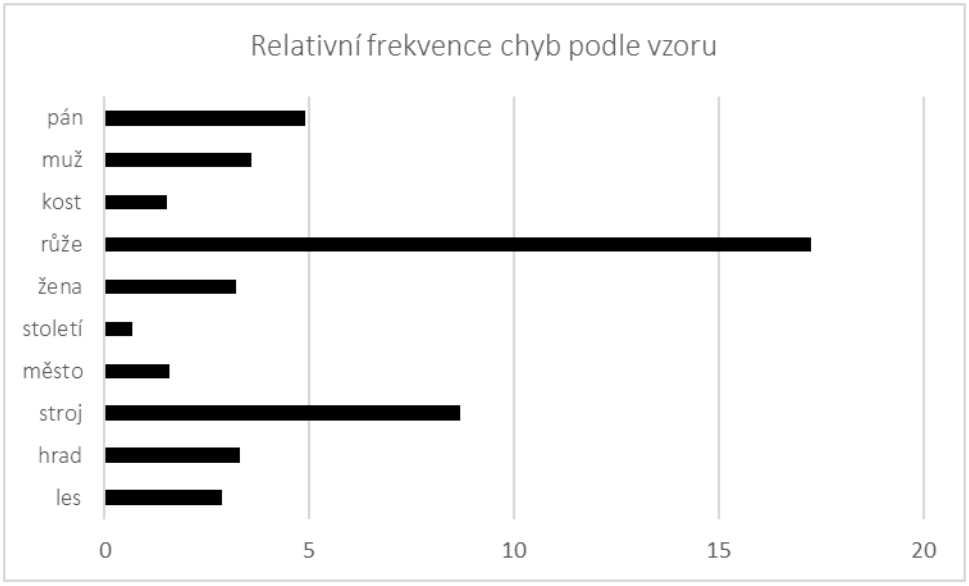
Tabulka 4.6

Pro představu o tom, jaké zkreslení by vniklo, pokud by relativní frekvence byly počítány podle tagů (tag1), které byly tvarům přiřazeny automaticky, uvádím následující tabulku.



Tabulka 4.7

Chyby v morfologické kategorii rodu se dají dále specifikovat na základě deklinačních vzorů, jak představuje **Tabulka 4.8**.[32]



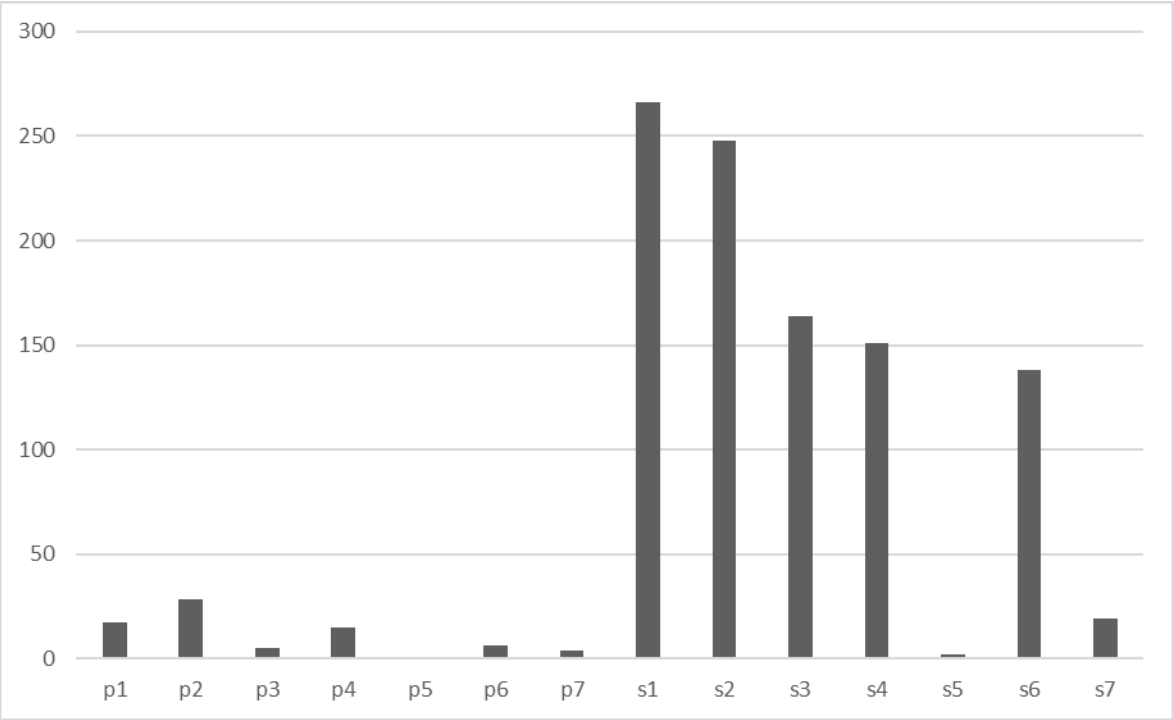
Tabulka 4.8

[32] Erratum: v tabulceTabulka 4.**8** má být místo „století“ uveden vzor „stavení“.

4.1.4 Rovina čtyři – zdroj chyby

Zatímco předchozí rovina analýzy specifikovala morfologické kategorie tvarů, ve kterých k chybám došlo, rovina 4 se zaměřuje na bližší morfologickou charakteristiku tvarů, na jejichž základě došlo k transferu.

K tvarům, které byly nejčastěji vybírány, patří nominativ a genitiv singuláru. Často je prostředkem chybného vyjádření také dativ, akuzativ a lokál singuláru. V grafu jsou zahrnuty všechny možné výskyty transferu na základě daných tvarů, tedy i kde se zároveň mohlo jednat o chybu v rodě, vzoru či čísle. Je třeba konstatovat, že počty jsou vychýlené častými případy pádové homonymie.



Tabulka 4.9 Zdroje transferu pro chyby v pádu

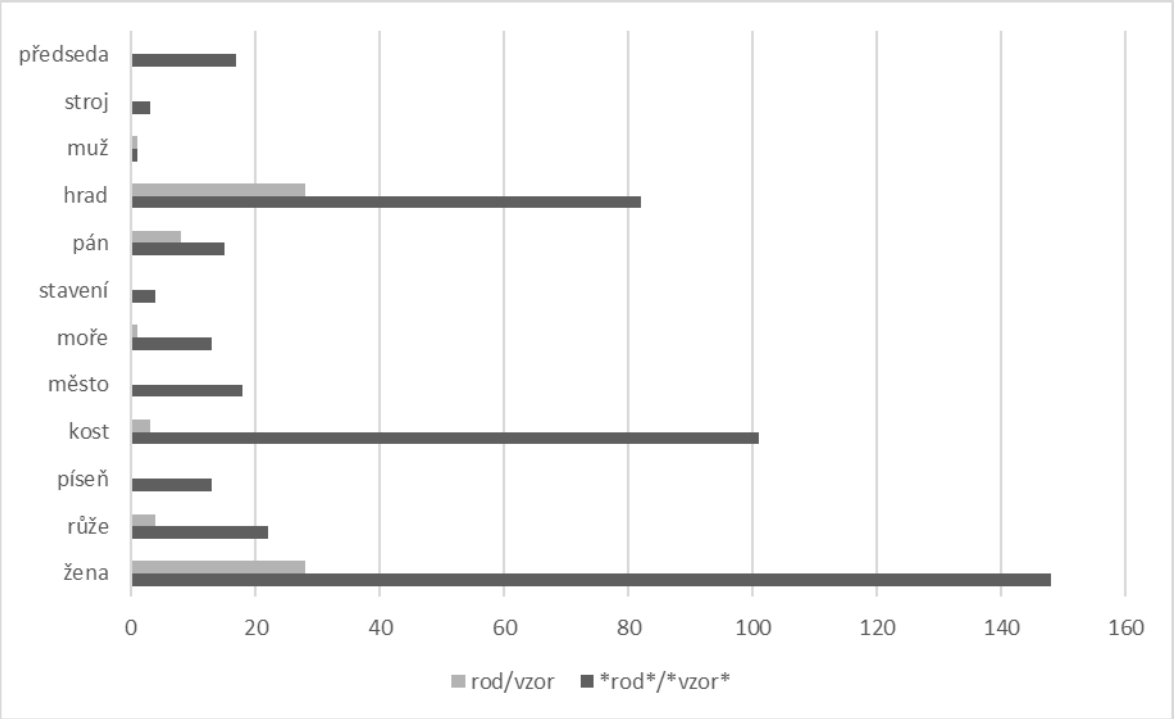
Pohled na nejčastěji transferované pády na větším počtu různých lemmat, který umožňuje korpus Merlin, ukazuje, že nerodilý mluvčí nominativní koncovku nejčastěji užívá v kontextech, které vyžadují akuzativ (v 30 ze 74 příkladů) nebo genitiv (14). Vyšší počet užití nominativu singuláru se objevuje i v plurálové formě genitivu (9), např.:

(44) Jedna z takových věc je myšlenka o tom, že víc hlav víc ví.

(45) Jeden z pozitivní strana žít v Praze je ten možnost cestovat.

Totéž platí i pro genitiv jako zdroj transferu, kde se k častějším tvarům, které přebírají genitivní koncovku, přidává ještě lokál singuláru.

Následující tabulka uvádí počet možných zdrojů transferu pro chybu v morfologickém rodu nebo deklinačním vzoru substantiv, černě pro všechny možné výskyty, šedě pro případy, kdy se jednalo jednoznačně pouze o skloňování podle příslušného rodu. Na tomto přehledu lze tedy sledovat míru specifičnosti určitých koncovek a vzorů. Například vzor předseda se jako možný zdroj transferu objevuje jen v případech, kdy jsou jeho deklinační zakončení totožná se vzorem žena. Jasně určit vliv vzoru na slova vzorů jiných tak lze v podstatě jen u vzoru hrad a žena.



Tabulka 4.10 Zdroje transferu pro chyby v rodu

4.1.5 Diskuze

Při srovnání výsledků tohoto výzkumu s pilotním výzkumem Schöne (2015) se jako nápadný jeví rozdíl v klasifikaci na rovině 1. Schöne jako pravopisnou chybu hodnotí jen 3,4 % (celkem

6) případů, zatímco zde tvoří většinu chyb 63,4 %. Schöne (2015: 25) navíc uvádí, že „ku prospěchu studenta byly chybné délky v morfologické koncovce posuzovány za chyby pravopisné“. Částečné vysvětlení nabízí možnost, že v jejím vzorku byly posuzovány pouze koncovky, nikoli například kořen slova, který se při ohýbání mění. Omezení pozornosti na pádové koncovky však v práci explicitně uvedeno není.

V ostatních aspektech a na ostatních rovinách vykazují zjištění této práce a pilotního výzkumu Schöne shodu. Rozložení morfologicky chybných tvarů podle bezprostřední složky si v obou vzorcích odpovídá. Na rovině 3 i přes drobné rozdíly v anotaci chybných morfologických kategorií se v obou případech výrazně projevuje tendence mnohonásobně častěji volit jiný pád než skloňovat substantivum v rámci jiného paradigmatu nebo zaměnit jeho číslo.

Výsledky obou analýz se shodují i při interpretaci zdroje transferu. Tedy ve vyhodnocení vzoru žena a hrad jako produktivních typů, podle nichž jsou skloňována substantiva příslušející v rámci stejného rodu k nedominantnímu vzoru i substantiva jiných rodů. Shoda platí i u chyb v pádech. Tvar, jímž jsou nejčastěji nahrazovány jiné pády, je nominativ singuláru. Častější užívání nominativu, než je běžné v korpusech rodilých mluvčích, není obecně pro žákovský jazyk nic neobvyklého. Vzhledem k jeho dominantnímu postavení jakožto reprezentanta celého paradigmatu a základního, výchozího tvaru, podle kterého se odvozují tvary ostatní, je velmi pravděpodobné, že se bude jednat o tvar nejsnadněji vybavitelný (srov. Schöne 2015: 30). Více než u vzorku Schöne, kde je druhým nejčastějším akuzativní transfer, se zde objevoval genitiv singuláru.

Na rozdíl od Schöne jsem se zde nezabývala tím, jak se četnost pádů transferu liší pro jednotlivé vzory a naopak. Schöne (2015: 31) konstatuje, že nominativ byl nejčastěji volen „bez zjevných rozdílů u jednotlivých deklinačních vzorů“. V následující části se ovšem zaměřím na to, jak se liší pro jednotlivé typy gramatických profilů.

4.2 Gramatické profily

V této části se budu zabývat tím, jaký vliv na produkci nerodilých mluvčích mají gramatické profily substantiv. Gramatický profil je k tomuto účelu chápán jako soubor dvou až tří nejfrekventovanějších tvarů, podle korpusu SYN2015 (Lehečková – Láznička – Janda 2016). Předpokládá se, že frekventovanější tvary budou produkovány procentuálně méně často než

tvary, které se nacházejí níže na frekvenční škále. Budu sledovat, jak se do tohoto předpokladu pomítají výše zjištěné strategie výpůjček deklinačních koncovek vzoru *žena* a *hrad* a nominativu singuláru.

	SYN2015			CzeSL-SGT		
lemma	1.	2.	3.	1.	2.	3
život	s2	s4	s6	s4	s1	s6
čas	s4	s2	s1	s4	s2	s1
svět	s2	s6	s1	s6	s1	s2
konec	s6	s2	s1	s6	s1	s2
případ	s6	p6	s4	s6	s1	p2
město	s2	s6	s1	s1	s6	s2
místo	s4	s6	s1	s1	p2	s4
slovo	s4	p2	p4	p2	p4	p1
jméno	s4	s7	s1	s1	s4	s7
století	s2	s6	p4	s2	s6	s4
práce	s2	s4	s1	s4	s2	s1
země	s6	s2	p2	s6	s1	p2
doba	s6	s4	s2	s6	s4	s2
strana	s6	s2	s4	s4	s6	s2
společnost	s2	s1	s6	s6	s1	s2
muž	s1	p1	p2	s1	p1	p2
otec	s1	s2	s7	s1	s4	s7
syn	s1	s4	s2	s1	s4	s7
autor	s1	p1	p2	s1	p1	p2
lékař	s1	p1	p2	s1	s3	p1

Tabulka 4.11 Nejfrekventovanější tvary v korpusu SYN2015 a CzeSL-SGT

4.2.1 Charakteristika vzorku

Nedá se předpokládat, že by se na tak malém vzorku (co do počtu lemmat různých rodů) příliš výrazně projevila tendence substantiv stejných profilů shlukovat se do skupin stejných sémantických či funkčních rysů (pro tento typ analýzy viz Schöne 2015), přesto lze tuto tendenci zřetelně pozorovat u maskulin, jež zahrnují rys životnosti.

Nejfrekventovanější pád všech životných maskulin ve vzorku je nominativ singuláru. V případě lemmatu *muž*, *autor* a *lékař* jako druhý nejfrekventovanější pád následuje nominativ plurálu a jako třetí genitiv plurálu. Schöne (2015: 126) ve svém výzkumu morfologických profilů v korpusu SYN2005 pozoruje převahu tohoto profilu (s1–p1–p2) v rámci sémantické třídy vymezené jako „člověk", podle jejího výzkumu je jedná profil typický pro názvy povolání, čemuž by odpovídat *lékař* i *autor*. Profil lemmat *syn* a *otec* se od profilu s1–p1–p2 ve druhé a třetí pozici z pochopitelných sémantických/pragmatických důvodů liší. Nominativ singuláru následují další singulárové tvary, které se spolu již neshodují (otec s1-s2-s7, syn s1-s4-s2).

Podobně zřetelně se životná maskulina klastrují i v projevech nerodilých mluvčích. Nejčastější tři tvary lemmat *muž*, *autor*, *lékař* odpovídají frekvenci v SYN2015 s jedinou výjimkou – v profilu lemma *lékař* se na druhé pozici objevuje dativ singuláru. Tato „odchylka" od korpusu rodilých mluvčích je poměrně snadno vysvětlitelná prominencí fráze *jít k doktorovi* / *jít k lékaři*. Například v učebnici Čeština expres (Holá – Bořilová 2015: 33) je dativ vykládán právě v lekci s tématem návštěvy u doktora. Zbylá dvě lemmata mužského životného rodu, u nichž jsou obvyklejší tvary singuláru, odpovídají stejnému profilu. Pro lemma otec i syn je po nominativu singuláru nejčastějším tvarem akuzativ singuláru a instrumentál singuláru (s1-s4-s7). Jsou to tvary, které dobře odpovídají jednoduchému popisu rodinných vztahů, které bývá ve výuce tematicky vyžadováno, např.:

(46) Můj otev má bratra v Tambovo . On má manželku a < syna > . Jeho manželka je učitelka a syn je malé díte .

(47) domácí úkoly . Dědeček a babička bydlí se svým mladším < synem > a jeho manželkou . Moje rodina je moc družná . Všichni

(48) mi 21 let . Jsem z Ruska z Petrohradu . Mám matku a < otce > . Otec je režisér a matka je letecký dispečer . Oba

(49) Eva Moje rodina . Má rodina neni velká . Bydlím s < otcem > a matkou . Jsem jeden syn v rodině . Ještě mám 2

Zbylá substantiva už tak snadno vymezitelná nejsou. Kromě nominativu singuláru je v SYN2015 u vybraných lexémů nejčastějším pádem genitiv singuláru (šestkrát), lokál singuláru (pětkrát) akuzativ singuláru (čtyřikrát). Klasifikace podle tří nejfrekventovanějších tvarů by vytvořila celkem 14 různých typů profilů (z celkem 15 lemmat), na základě dvou nejfrekventovanějších by vzniklo 10 typů a bez ohledu na pořadí první tří nejfrekventovanějších tvarů by jich bylo 8, při omezení jen na dva tvary bez ohledu na pořadí 7.

Tak velká diverzita není u malého vzorku nijak neobvyklá. Z typů, které jsou v něm zastoupeny více než jedním lemmatem, lze ještě vybrat profil s2-s6-s1. Náleží k němu lemmata *město* a *svět.* Schöne typ s2-s6, ke kterému tedy řadí i podtyp s2-s6-s4 (v tomto vzorku lemma *století*) charakterizuje jako substantiva s primárně lokálním výrazem, která se typicky vyskytují s předložkami místa a směru *v*, *do* a *na* (Schöne 2015: 120). Tyto tvary patří u v korpusu CzeSL pro substantiva *město*, *svět* a *století* k těm nejfrekventovanějším, ačkoli se plně neshodují v pořadí tvarů. Například nejvyšší frekvence nominativu u lemmatu *město* místo genitivu, který pokrývá nejvíce výskytů města v korpusu SYN2015, je ovšem snadno vysvětlitelná pomocí metadat korpusu CZeSL. Ta ukazují, že poměrně velká část konkordance města vznikla jako text na téma *Mé rodné město* / *Město, ve kterém bych chtěla žít*. V korpusu pak nejen, že se tato konstrukce často opakuje, ale určuje u konstrukce další, např.:

(50) Ahoj Anton ! Desatého šestého listopadu . Mé rodné < město > se jmenuje Dimitrovgrad . Město se nachází vědle

(51) Cestovala bych hodně , objela bych všechna města . < Město > v kterém chci žít . Mé rodné a milované město je

(52) dobře . Město , ve kterem chtěl bych žít Myslím že < město > , ve kterem chtěl bych žít , musí být velké , hezké

Schöne (2015) při konstruování morfologických profilů odhlíží od frekvenčního pořadí nejčetnějších tvarů, zajímá jí tedy, jaké dva (případně tři) tvary se vykytují nejčastěji bez ohledu

na to, který z nich je první a který druhý. Při takovém postupu by k typu s2-s6-(*) bylo možné přiřadit též lemmata *země* (s6-s2-p2), *konec* (s6-s2-s1) a *strana* (s6-s2-s4), všechna sémanticky zapadají do výše popsaného profilu. V těchto případech se frekvence lemmat v korpusu nerodilých mluvčích různí o něco víc. Pro *konec* a *zemi* je sice nejfrekventovanější lokál singuláru, genitiv singuláru se ovšem u lemmatu *konec* vyskytuje na třetí a u lemmatu *země* až na čtvrté pozici. Četnější je v obou případech nominativ singuláru a pro lemma *země* též i genitiv plurálu. Lemma *strana* se od vzorce odchyluje tím, že nejčastěji se vyskytuje v akuzativu singuláru, a lokál s genitivem tudíž až na druhé a třetí pozici. Toto vychýlení je jednoznačně způsobené užíváním párové fráze *na jednu stranu – na druhou stranu*, které se v žákovských projevech často až mechanicky opakuje, např.:

(53) Mám Českou republiku velmi ráda , proto že na jednou < stranu > tady bydlí čast mé rodiny . Na druhou < stranu > miluju český jazyk že pracování s českým

Častější výskyt nominativu je ve shodě s výše uvedeným předpokladem jeho didaktického významu. Nadto opět částečně vyvoláno tématem textů – *Proč mám/nemám rád ČR* / *Moje země a ČR*, např:

(54) se oživají slavnostníy a tradicí . ČR je kulturní < země > ale navíc je i pro děti : Rodiče mají možnost posilat

(55) republiku Původem jsem z Moldávie . Je to malá a chudá < země > mezi Ukrajinou a Rumunskem . Do svých 11 let jsem

(56) vetšinou ostatních evropských zemí , Česko je turistická < země > , tady hodně turistů je navždy na ulicích , tady

K dalšímu typu by se dala přiřadit lemmata *život* a *práce* (s2-s4), bez přihlédnutí k frekvenčnímu pořadí též lemma *čas*. V současné češtině se jedná o nejběžnější, proto také komplikovaně vymezitelný, typ frekvenčního seskupení tvarů (Schöne 2015: 112). V relativně vyrovnaném počtu obsahuje jak jména abstraktivní, tak konkrétní, jak nepočitatelná, tak počitatelná (tamtéž), skutečnost, že se do tohoto typu vyprofilovala výše uvedená substantiva tedy pravděpodobně neodráží jeho charakteristiku. V korpusu CzeSL je ve všech třech případech nejčastějším tvarem akuzativ singuláru, u lemmatu *práce* a lemmatu *čas*, jejichž

profily se zcela shodují, následuje genitiv singuláru a na třetí pozici nominativ singuláru. U lemmatu život je genitiv singuláru až na čtvrtém místě, předchází ho akuzativ (na prvním místě), nominativ a lokál. Preference nominativu je vysvětlitelná výše popsanými důvody, ještě více než v předešlých případech ovšem platí ovlivněnost této formy zadáním. Celá řada témat o různých formách života tentokrát obsahuje nejen nominativ, ale i akuzativ a lokál, čímž by se vychýlení dalo vysvětlit celé – *Člověk, který mi v životě nejvíc pomohl* / *Jakou barvu má život?* / *Život v cizině* / *Jak bude vypadat můj život za pět let* / *Život na koleji vs. život doma* / *Událost, která změnila můj život* / *Řídí můj život média?* / *Je život bez rodiny těžký?* Zdravý život (ovšem též *Škola základ života*), např.:

(57) Jakou barvu má < život > ? Kdyby < život > *měl barvu* , byl by . . . To opravdu nevím . Kdy < život > plynuje , mění se , nezamýšleme se o něm . < Život > sám po sobě je zelený . Jako každá rostlina má na začátku zelenou barvu , tak i < život > .

(58) Člověk , *který mi v životě* nejvíc pomohl to je můj otec . Jimenuje se Adam . mají šedive kratke vlasy , zelene ocí a trochu tlustý . Cely život se stara o svou rodinu . Když byl mlády pracoval v lodi a mnel velké přaní

(59) Člověk , *který mi v životě* nejvíc pomohl . Myslím si , že se každí rodiče pořád starají o svou děti a nejvíc jim pomůžou . Moje rodiče mi taky nejvíc pomohli v životě , vlastně maminka . Maminka je mílá a starostlivá , ale někdy

Kromě typu s2-s6 (viz výše) vymezuje Schöne jako profil typický pro substantiva s lokálním i několik substantiv s časovým významem typ s6-s4 (2015: 124). K tomuto profilu lze přiřadit lemma *doba* (s6-s4-s2) a lemma *místo* (s6-s4-s1). Zatímco profil *doby* se v korpusu SYN a korpusu CzeSL shodují ve všech třech sledovaných tvarech, *místo* představuje další „odchylku“. Ze tří v korpusu SYN nejfrekventovanějších tvarů obsahuje jeho reprezentace v CzeSL jen akuzativ singuláru a to na třetí pozici. Častější je nominativ singuláru a poměrně překvapivě také genitiv plurálu.

Lemma *společnost* lze považovat za zástupce poměrně frekventovaného, avšak těžko charakterizovaného typu s2-s1. S lokálem singuláru na třetí pozici navíc společnost

pravděpodobně představuje málo frekventovaný podtyp (Schöne 2015: 123). Až na posun v pořadí lokálu singuláru (na první místo) jeho profil v CzeSL odpovídá výsledkům ze SYN2015.

Samostatné, a z hlediska relativní frekvence ve větším vzorku (Schöne 2015: 112) ne příliš časné profily tvoří lemmata *jméno* (s4-s7-s1), *slovo* (s4-p2-p4). Instrumentál singuláru a genitiv plurálu na druhé pozici, které je činí specifickými, i na první pozici běžný akuzativ singuláru, obsahují i jejich pandány z korpusu nerodilých mluvčí, ačkoli ne ve stejném pořadí. Lemma jméno odpovídá i v dalším tvaru – nominativu singuláru, pro lemma slovo je na třetí pozici v CzeSL častější nominativ plurálu než akuzativ plurálu tak, tak je tomu v SYN2015.

I přes určité posuny v pořadí lze konstatovat, že vybraná lemmata mají v korpusu nerodilých mluvčí tendenci kopírovat frekvenční charakteristiky, které daná slova mají v reprezentativním korpusu současné češtiny. Případy, kdy tomu tak není lze většinou vysvětlit za pomoci metadat, která korpus CzeSL obsahuje. Na druhou stranu například pro genitiv plurálu jako druhého nejčetnějšího tvaru slova *místo*, se příliš vysvětlení nenabízí. Podle popsaných skupin či jednotlivých typů profilů je tedy lemmata možné roztřídit do následujících skupin.

				SYN2015			CzeSL-SGT		
			lemma	1.	2.	3.	1.	2.	3.
	s1	s1-p1-p2	muž	s1	p1	p2	s1	p1	p2
			autor	s1	p1	p2	s1	p1	p2
			lékař	s1	p1	p2	s1	**s3**	p1
			otec	s1	s2	s7	s1	**s4**	s7
			syn	s1	s4	s2	s1	s4	**s7**
	s2-s6	s2-s6-s1	svět	s2	s6	s1	s6	s1	s2
			město	s2	s6	s1	s1	s6	s2
			století	s2	s6	p4	s2	s6	**s**4
			konec	s6	s2	s1	s6	s1	s2
			země	s6	s2	p2	s6	**s1**	p2
			strana	s6	s2	s4	s4	s6	s2
			život	s2	s4	s6	s4	**s1**	s6
			čas	s4	s2	s1	s4	s2	s1
			práce	s2	s4	s1	s4	s2	s1
			místo	s4	s6	s1	s1	**p2**	s4
			doba	s6	s4	s2	s6	s4	s2
			společnost	s2	s1	s6	s6	s1	s2
			jméno	s4	s7	s1	s1	s4	s7
			slovo	s4	p2	p4	p2	p4	**p1**
			případ	s6	p6	s4	s6	**s1**	**p2**

Tabulka 4.12 Gramatické profily podle typu (na základě SYN2015)[33]

4.2.2 Chybovost podle profilů

Mají-li gramatické profily vliv na produkci nerodilých mluvčích, lze (kromě výše analyzovaného odrazu frekvenčních charakteristik) také předpokládat jednak, že u substantiv stejných frekvenčních typů bude existovat podobnost i s ohledem na tvary, v nichž nerodilý mluvčí nejčastěji chybují, a jednak to, že tvary, které tvoří profil substantiva (nejfrekventovanější tři či dva), budou produkovány s nejnižší chybovostí. Tabulka 4.13 uvádí rozložení chyb podle tvarů a jejich relativní frekvenci vzhledem k celkovému počtu výskytu daných tvarů v CzeSL (podle opravených tagů (tag2)):

[33] Zvýrazněné jsou tvary, které se mezi třemi nejfrekventovanějšími objevují pouze v korpusu CzeSL.

	muž	autor	lékař	otec	syn	svět	město	století	konec	země	strana	život	čas	práce	místo	doba	spol.	jméno	slovo	případ
p1	10	1	0	0	0	0	0	0	0	6	0	0	0	0	1	0	0	0	0	0
p2	9	0	0	0	0	0	0	0	0	15	0	0	2	7	4	0	0	0	3	0
p3	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
p4	9	0	1	0	7	0	4	0	0	7	0	2	0	2	5	0	0	0	3	1
p5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p6	2	0	0	0	0	0	6	0	0	21	2	0	2	0	1	1	0	0	0	0
p7	0	0	0	0	0	0	2	1	0	8	1	0	0	0	0	0	0	0	1	0
s1	3	1	1	0	3	0	4	0	0	6	0	0	0	5	0	0	0	0	3	0
s2	1	1	1	0	0	24	13	0	3	12	0	46	23	66	1	0	0	0	1	0
s3	1	0	4	0	0	0	1	0	1	0	0	1	0	5	0	0	0	0	0	0
s4	11	1	0	0	4	1	4	0	1	8	20	5	22	106	4	4	2	0	1	0
s5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s6	2	0	0	0	0	13	11	0	18	61	0	16	5	19	2	4	1	0	1	0
s7	2	0	0	0	0	0	2	0	1	1	0	3	2	0	0	0	0	0	0	0
	muž	autor	lékař	otec	syn	svět	město	století	konec	země	strana	život	čas	práce	místo	doba	spol.	jméno	slovo	případ
p1	4,7	16,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	25,0	0,0	0,0	0,0	0,0	2,1	0,0	0,0	0,0	0,0	0,0
p2	25,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	10,4	0,0	0,0	9,5	25,9	2,3	0,0	0,0	0,0	2,6	0,0
p3	0,0	0,0	100	0,0	100	0,0	20,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	20,0	0,0
p4	25,7	0,0	50,0	0,0	58,3	0,0	5,9	0,0	0,0	13,5	0,0	16,7	0,0	3,6	6,1	0,0	0,0	0,0	4,0	16,7
p5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
p6	50,0	0,0	0,0	0,0	0,0	0,0	9,1	0,0	0,0	61,8	0,0	0,0	100	0,0	4,5	16,7	0,0	0,0	0,0	0,0
p7	0,0	0,0	0,0	0,0	0,0	0,0	33,3	33,3	0,0	80,0	50,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	4,5	0,0
s1	0,5	1,9	2,8	0,0	1,7	0,0	0,3	0,0	0,0	3,3	0,0	0,0	0,0	3,4	0,0	0,0	0,0	0,0	7,1	0,0
s2	6,7	16,7	33,3	0,0	20,0	17,5	3,0	0,0	9,7	10,6	0,0	8,7	6,3	25,3	2,4	0,0	0,0	0,0	4,5	0,0
s3	9,1	0,0	26,7	0,0	0,0	0,0	10,0	0,0	12,5	0,0	0,0	1,9	0,0	17,9	0,0	0,0	0,0	0,0	0,0	0,0
s4	30,6	33,3	0,0	0,0	4,5	1,3	2,1	0,0	4,8	15,4	38,5	0,5	2,5	26,6	3,1	2,6	8,3	0,0	2,4	0,0
s5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
s6	25,0	0,0	0,0	0,0	0,0	2,7	1,8	0,0	20,5	15,8	0,0	2,4	10,4	14,4	0,0	1,3	1,6	0,0	16,7	0,0
s7	9,5	0,0	0,0	0,0	0,0	0,0	2,2	0,0	16,7	10,0	0,0	6,0	7,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Tabulka 4.13 Absolutní a relativní frekvence chyb podle lemmat

V tabulceTabulka 4.14 pro každé lemma vybrány tři nejčastěji chybné tvary, jak v absolutních, tak v relativních číslech. Lemmata, která zahrnovala pouze tvary, v nichž se chyba vyskytla jedenkrát (*autor*, *století*, *případ*), popřípadě se v nich morfologicky chybné tvary neobjevily vůbec (*otec*, *jméno*), nebyla do analýzy zahrnuta.

lemma	1.	2.	3.	1.	2.	3.	1.	2.	3.	1.	2.	3.
muž	s1	p1	p2	s1	p1	p2	s4	p1	p2/p4	p6	s4	p4
~~autor~~	s1	p1	p2	s1	p1	p2						
lékař	s1	p1	p2	s1	**s3**	p1	s3	s1/s2/p3/p4	s1/s2/p3/p4	p3	p4	s2
~~otec~~	s1	s2	s7	s1	**s4**	s7						
syn	s1	s4	s2	s1	s4	**s7**	p4	s4	s1	p3	p4	s4
svět	s2	s6	s1	s6	s1	s2	s2	s6		s2	s6	
město	s2	s6	s1	s1	s6	s2	s2	s6	p6	p7	p3	s3
~~století~~	s2	s6	p4	s2	s6	**s4**						
konec	s6	s2	s1	s6	s1	s2	s6	s2		s6	s7	s3
země	s6	s2	p2	s6	**s1**	p2	s6	p6	p2	p7	p6	p1
strana	s6	s2	s4	s4	s6	s2	s4	p6		s4	p7	
život	s2	s4	s6	s4	**s1**	s6	s2	s6	s4	p4	s2	s7
čas	s4	s2	s1	s4	s2	s1	s2	s4	s6	p6	s6	p2
práce	s2	s4	s1	s4	s2	s1	s4	s2	s6	p4	p6	s4
místo	s4	s6	s1	s1	**p2**	s4	p4	p2/s4	p2/p4	p4	p6	s4
doba	s6	s4	s2	s6	s4	s2	s6/s4	s6/s4	p6	p6	s4	s6
společnost	s2	s1	s6	s6	s1	s2	s4			s4		
~~jméno~~	s4	s7	s1	s1	s4	s7						
slovo	s4	p2	p4	p2	p4	**p1**	s1/p2/p4	s1/p2/p4	s1/p2/p4	p3	s6	s1
~~případ~~	s6	p6	s4	s6	**s1**	**p2**						

Tabulka 4.14 Absolutní a relativní frekvence chyb podle profilů

Z absolutního a relativního porovnání nejčastějších chyb je zřejmé, že absolutní počet chyb je spíše ukazatelem frekvence daného tvaru, než pravděpodobnosti, že nerodilý mluvčí udělá chybu právě v tomto tvaru.

V případě substantiv typu s1-p1-p2 připadají tři relativně nejčastější chyby na tvary, z nichž ani jeden netvoří gramatický profil substantiva. Tyto chyby lze interpretovat buď jako tendenci k užití častějšího pádu, nebo jako tendenci ke skloňování podle měkkých ženských vzorů, např.:

(60) skutečnosti ženy budují svou kariéru a jsou nezavisle od < muži > . Tak myslím , že přes 50 let rozdíl mezi podnikátélem

(61) Každý den měří teplotu a artériový tlak , znají všech < lékařů > ve své nemocnici a když se necítí dobře (může byt

Podobně se chová i substantivum *syn*. Jako třetí relativně „nejchybovější“ pád se ovšem objevuje akuzativ singuláru, který je druhým nejčastějším tvarem. Tyto výskyty lze ve většině případů hodnotit buď jako upřednostnění frekventovanějšího nominativu, nebo příklon k neživotným substantivům, např.:

(62) smutno protože jsou velmi osamělé . Tento rok jejich < syn > a dceru přestěhovali do Ameriky Prahy v letě . Tak

(63) je dcera . Protože jsem trochu mám ráda dcera než < syn > . Doufám moje děti jsou sympatické a zdravé . Když

(64) Evu . Můj strýc Adam má manželku Evu , oni mají 3 < synové > : Adama , Adama a Adama . Adam je starší , on na

I okolí KWIC v uvedených případech ukazuje, že je velmi komplikované zhodnotit, která hypotéza (nominativ vs. neživotnost) je pravděpodobnější. Příklady (63) a (64) ukazují spíše na vliv nominativu, neboť v (63) stojí v akuzativu též substantivum dcera a v (64) je užita koncovka životných maskulin, příklad (62) a druhý člen parataktického spojení zase mluví ve prospěch hypotézy o vlivu neživotného vzoru.

V případě skupiny s profilem s2-s6/s6-s2 vymezitelné jako substantiva lokálního významu (*svět*, *město*, *konec*, *země*, *strana*) se tyto dva nejfrekventovanější pády objevují jen lemmatu *svět* (singulár i lokál) a u lemmatu *konec*. Oběma těmto tvarům ovšem v CzeSLu schází prakticky celé plurálové paradigma. Svou roli dále sehrává příslušnost obou tvarů k méně typickým (pod)vzorům neživotných maskulin (*stroj* a *les*), jak lze pro lokál singuláru demonstrovat na následujících příkladech:

(65) s různými staty a také vědem co se noveho stalo ve < světu > . Potřebujeme média aby vědeli jestli něco stalo

(66) ješte studuju doma . Chci vyborně udelat zkoušky na < koncu > semestra . Večer poslouhám rádio Jedná nebo se divám

Další chyby v lokálu singuláru jsou hodnoceny jako chyby v pádu (je užit tvar nominativu/akuzativu), případný tvar akuzativu navíc souzní s možnou valencí předložky na:

(67) vzduch . Pro mě ty bázni jsou nejkrásnější věci na < svět > , bez kterých nechci žít . Myslím , že mám nejbližši

(68) také s lidmi , kteří neumí ruský či anglický . Na < konec / > listopadu mohl jsem už mluvit pořad bez ohledu na

V případě genitivu singuláru, jakožto nejfrekventovanějšího tvaru lemma *svět* i *tvaru*, který je produkován s nejvyšší chybovostí, platí, že příklonem ke koncovce slova *hrad*, lze téměř vysvětlit všechny chyby. Ve dvou případech (s2 a s4) lze narazit i na přechod k femininu a skloňování podle vzoru žena, např.:

(69) Švýcarsku často mluví , že ten stát je hodinková fabrika < světu > . Hlavní pamětihodnost jsou čokoláda a sýr , které

(70) Švédska na jaře , protože tam se bude konat Misterstvo < světu > . Kdybych měla dost peněz , jela bych na Olimpické

(71) on chce být mladým . Také líbí se mi cestovat kolem < světu > . Mýslím že v budoucnosti budu dělat to hodně . Také

(72) protože si myslím , že prážský hrad je nejhezčí z celé < světy > , a proto ho musíme uvidět spolu , přestože už jsi

Z ostatních morfologicky chybných tvarů, které odpovídají žádnému profilu substantiva této skupiny, se nejčastěji objevuje instrumentál plurálu. Ten se zároveň mezi třemi nejčastěji špatně zformovanými nebo zvolenými pády objevuje výhradně v této skupině. Příznačně i u lemmatu *století*, které bylo vyřazeno z analýzy, protože obsahuje jediný chybný tvar, je jedná o instrumentál lokálu. Např.:

(1) druru se patří vylety např. do Evropy . Nejznamějšími < zemí > pro cestování jsou : Nemecko , Italie , Česká Republika

(2) jazyk trochu stejný a ruštinou . Vím že před nekolika < století > naší narody měli jeden jazyk a proto mají hodně společného

(3) dům a moje rodina . Kišiněv má zvláštní status mezi < měst > Moldavska , je municiXXXni co znaméná co 6 měst co

(4) vídí příčiny této události v nesouhlasu mezi dvěma < strány Quant1 > a nekompetenci ruské autority a vlády . Podle mého

U dvou lemmat se ještě frekventovaně vykytuje chyba v dativu singuláru, což je také specifikum pouze této skupiny. Všechny ostatní chyby časté pro substantiva typu s2-s6/s6-s2 jsou rozloženy do různých tvarů (s4, s7, p1, p3, p6)

U substantiv typu s2-s4/s4-s2 jsou nejčastěji chybnými tvary akuzativ a lokál plurálu. Jako nejčastěji chybný tvar je dvakrát zastoupen i tvar nejčastěji se vyskytující (s2 na druhé pozici lemmatu *život* a s4 na třetí pozici lemmatu *práce*). V zásadě totéž platí i pro lemma *doba* a *místo*, i v případě těchto substantiv spočívají nejčetnější chyby v plurálových tvarech akuzativu a lokálu a na druhé a třetí pozici se objevují frekventované tvary. Lemma společnost (typ s2-s1-s6) nejčastěji chybuje v akuzativu singuláru. Specifický vzorec tvoří pořadí chyb lemmata *slovo*. Mezi tři nejčastější chyby se dostává nominativ singuláru.

5 Závěry

Tato práce se zabývala frekvenčními charakteristikami českých substantiv a jejich odrazem v produkci nerodilých mluvčích. Snažila se přitom zodpovědět otázku, zda lze na míře chybovosti pozorovat frekvenční efekt gramatických profilů substantiv a jakými dalšími faktory lze chyby, kterých se mluvčí dopouštějí nejčastěji, vysvětlit.

Analytická část ukázala, že v produkci nerodilých mluvčích lze pozorovat tendenci kopírovat frekvenční chování vybraných lemmat a do jisté míry též vliv těchto frekvenčních profilů na distribuci chyb. I v rámci relativně malého vzorku 20 substantiv bylo možné na základě dvou až tří nejfrekventovanějších tvarů vytvořit několik substantivních klastrů vykazujících podobné rysy – třídu s1-p1-s1 (obsahující lemmata *muž*, *autor*, *lékař*); třídu s1-s4-s7 (*otec*, *syn*); s2-s6/s6-s2 (*svět*, *město*, *století*, *konec*, *země*, *strana*); s2-s4/s4-s2 (život, čas, práce); s4-s6/s6-s4 (místo a doba); zbylá lemmata (*společnost*, *jméno*, *slovo*, *případ*) tvořila samostatné typy. Jen ve dvou případech se nejčastější pád objevuje i nejčastěji chybně. Nadto se zdá, že nejen, že je možné podle gramatického profilu substantiva odhadovat, jaké chyba bude méně častá, platí to i naopak – podle chyby je někdy možné usuzovat na profil substantiva, určité tvary se jako nejčastěji chybné objevují jen v rámci určitého klastru (např. instrumentál plurálu třídy s2-s6/s6-s2). Dala by se rozsáhlejším vzorku podle distribuce chyb seskupit do stejného klastru jako podle nejčetnějších pádů?

Případy, v nichž výše uvedené neplatí, tedy kdy se nejčastější tvar objevuje jako častěji chybný, lze hodnotit jako doklady efektu produktivního typu. (Například u lemmatu svět nejčastěji chybné tvary korespondují s tvary nejčastěji se vyskytujícími, tj. s2 a s6. Právě v těchto tvarech se projevuje rozdíl mezi koncovkami vzoru hrad a jeho podvzoru les.) Rozhodně to tak není vždy a efekt produktivity konstrukce nelze považovat za „silnější" efekt, ani za efekt jediný působící jiným směrem (jako příklad může sloužit tvar lokálu singuláru u lemmat země a doba. V obou případech se jedná o nejfrekventovanější tvar, u lemmatu *země* by se dala čekat tendence přiklonit se ke vzoru žena, lokál singuláru se mezi nejčastěji chybnými ovšem vyskytuje pouze u lemmatu *doba* – místo lokálu je volen akuzativ (*Žijeme v < dobu > informačních technologií*)). Jaké další faktory je třeba zohlednit?

Analýza rozložení chyb podle bezprostředních složek potvrdila, že nejvíce chyb připadá na předložkové fráze fungující jako příslovečná určení místa nebo směru (srov. Schöne 2015). Analýza nejčastějších chyb zároveň ukázala, že v případech, kdy neplatí, že nejčastější tvar substantiva nebude mezi tvary nejčastěji chybnými, jedná se často o přeložkové pády. Tvary po předložkách abstraktnějšího významu tvořící součást verbální fráze jsou produkovány s nižší chybovostí. To může být dáno reprezentací frází druhého typu, nebo například i tím, že v kontextu slovesa pohybu a předložky může být indikace pádu koncovkou pociťována jako méně důležitá / k porozumění méně nutná (viz efekt redundance). Takovou hypotézu by bylo třeba dále ověřit například psycholingvistickým experimentem nebo podrobnějším porovnáním chybně užitých substantiv určujících místo ve verbálních vs. nominálních frázích na rozsáhlejším korpusovém vzorku.

Výše uvedené otázky představují příklad směru, kterým by se mohl ubírat budoucí výzkum. Výše uvedené závěry narážejí na zásadní problém. Tím je otázka, nakolik můžeme důvěřovat datům akvizičních korpusů, resp. nakolik je lze zobecnit. I přesto, že výzkumným vzorkem této práce byla nejfrekventovanější substantiva současné češtiny, pozorované tendence se často zakládají pouze na několika příkladech. Toto omezení je třeba mít na paměti. Bývá také dobrým zvykem, že ručně anotované texty procházejí vícenásobnou kontrolou, než se jako korpusová data zpřístupní badatelům. Pro účely této práce jsem anotace zhruba 20 000 tvarů z kapacitních důvodů nekontrolovala. Vzhledem k nízkému počtu výskytů může mít i ojedinělá chyba v anotaci zásadní důsledky pro platnost závěrů.

Tato práce vzniká v určitém mezidobí, kdy i pro češtinu vznikají poměrně rozsáhlé žákovské a akviziční korpusy, problematikou, která je aktuální, je však právě jejich vytváření, k němuž existuje řada studií (např. Šebesta – Škodová 2012; Štindlová – Čurdová 2015; Štindlová a kol. 2014), více stranou zůstává otázka jejich vytěžování. Otázku po tom, zda z nich lze čerpat obecně(ji) platné informace o (mezi)jazyce nerodilých mluvčí, tak bude pro češtinu pravděpodobně možné zodpovědět až po zpřístupnění ručně anotované verze korpusu CzeSL.

6 Bibliografie

Boyd, Adriane et al. (2014): The MERLIN corpus: Learner Language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), European Language Resources Association (ELRA), Reykjavik, May 26-31, 2014.

BYBEE, Joan (2006): From usage to grammar: the mind's response to repetition. Language, 82, 711–733.

BYBEE, Joan (2008): Usage-based grammar and second language acquisition. In: P. Robinson – N.C. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquistion*. New York: Routledge, 216–236.

CVRČEK, Václav a kol. (2010): *Mluvnice současné češtiny 1: Jak se píše a jak se mluví*. Praha: Karolinum.

CVRČEK, Václav – RICHTEROVÁ, Olga (eds.) (2013): *Příručka ČNK* [on-line], pojmy: frekvence; CzeSL-plain; CzeSL-SGT; SYN2015. [cit. 24-07-2016]. Dostupné z WWW: https://wiki.korpus.cz/doku.php?id=pojmy:frekvence&rev=1379494352.

DIVJAK, Dagmar – CALDWELL-HARRIS, Catherine (2015): Frequency and entrenchment. In: Ewa Dąbrowska – Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*. Berlin: De Gruyter, 53–75.

DIVJAK, Dagmar – GRIES, Stefan Th. (2008): Clusters in the mind? Converging evidence from near synonymy in Russian. *Mental Lexicon*, 3, 188–213.

DIVJAK, Dagmar – GRIES, Stefan Th. (2006): Ways of trying in Russian: Clustering. *Corpus Linguistics and Linguistic Theory*, 2, 23–60.

ECKHOFF, Hanne M. – JANDA, Laura A. (2014): Grammatical profiles and aspect in Old Church Slavonic. *Transactions of the Philological Society*, 112, 231–258.

GASS, Susan M. – MACKEY, Alison (2002): Frequency effects and Second Language Acquisition. A complex picture? Studies in Second Language Acquisition, 24, 249–260

GRIES, Stefan Th. (2011): Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In: Mario Brdar, Milena Žic Fuchs – Stefan Th. Gries (eds.), *Expanding Cognitive Linguistic Horizons*. Amsterdam – Philadelphia: John Benjamins.

GRIES, Stefan Th. – DIVJAK, Dagmar (2009): Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: Vyvyan Evans – Stephanie Pourcel (eds.), *New Directions in Cognitive Linguistics*. Amsterdam – Philadelphia: John Benjamins, 57–75.

HOLÁ, Lída – BOŘILOVÁ Pavla (2015): Čeština expres 2: A1/2. Praha: Akropolis.

HOLÁ, Lída: Dá se čeština číst odzadu? Poznámka k výuce češtiny pro cizince [on-line, cit. 15-08-2016].

ELLIS, Nick C. (2002): Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. Studies in Second Language Acquisition, 24,143–188.

ELLIS, Nick C. (2014): Frequency-based accounts of Second Language Acquisition. In: S.M. Gass – A. Mackey, The Routledge Handbook of Second Language Acquisition. New York: Routledge, 2014, 193–210.

ELLIS, Nick C. – WULFF, Stefanie (2015): Second language acquisition. In: Ewa Dąbrowska – Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*. Berlin: De Gruyter, 409–432.

JANDA, Laura A. – LYASHEVSKAYA, Olga (2011): Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. *Cognitive Linguistics*, 22, 719–763.

KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.) (2002): *Encyklopedický slovník češtiny*. Praha: NLN.

KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.) (v tisku): *Nový encyklopedický slovník češtiny*. Praha: NLN.

KŘIVAN, Jan (2012): Komparativ v korpusu: explanace morfematické struktury českého stupňování na základě frekvence tvarů. *Slovo a slovesnost*, 73, 13–45.

LEHEČKOVÁ, Eva – LÁZNIČKA, Michal – JANDA, Vojtěch (2016): Frequency-based grammatical profiles of Czech nouns. Poster na konferenci The 2nd usage-based linguistics conference. Tel Aviv and Hebrew University.

MERLIN PROJECT. 2014. *Merlin* [Online]. 2014 [cit. 27.05.2016], Annotation scheme. Dostupný z WWW: <http://commul.eurac.edu/dev/merlin/php/docs/MERLIN-annotation-scheme.pdf>.

PFÄNDER, Stefan – BEHRENS, Heike (2016): Experience counts: An introduction to frequency effects in language. In: Heike BEHRENS – Stefan PFÄNDER (eds.), *Experience Counts: Frequency Effects in Language*. Berlin – Boston: De Gruyter.

ROSEN, Alexandr (2015): CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací. Ústav teoretické a komputační lingvistiky [online, cit. 29-07-2016]. Dostupné z WWW: http://utkl.ff.cuni.cz/~rosen/public/2014.

SCHÖNE, Karin (2015): Zkoumání hierarchizace pádů českého substantiva v sémantických (kolokačních) třídách. Nepublikovaná disertační práce. Praha: FF UK.

SCHÖNE, Karin (2011): Úvahy o prezentaci substantivního skloňování v učebnicích češtiny z frekvenčního hlediska. In: F. Čermák (ed.), Korpusová lingvistika Praha 2011. Praha: Nakladatelství Lidové noviny.

ŠEBESTA, Karel et al. (2014): *CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014*. Praha: Ústav Českého národního korpusu FF UK. Dostupné z WWW: http://www.korpus.cz.

ŠEBESTA, Karel (2012): Parametry žákovských korpusů a CzeSL. In: K. Šebesta – S. Škodová (eds.), *Čeština – cílový jazyk a korpusy*. Liberec: Technická univerzita v Liberci.

ŠEBESTA, Karel – ŠKODOVÁ, Svatava (2012) (eds.): *Čeština – cílový jazyk a korpusy*. Liberec: Technická univerzita v Liberci.

ŠTINDLOVÁ, Barbora – ČURDOVÁ, Veronika (2015): Merlin: Multilingvální platforma pro evropské referenční úrovně. *Časopis pro moderní filologii*, 97(2). Praha, 190–200.

ŠTINDLOVÁ, Barbora a kol. (2014): Žákovský korpus Merlin: jazykové úrovně a trojjazyčná chybová anotace. In: *Práce s chybou ve výuce cizích jazyků (včetně češtiny pro cizince)*. Sborník z mezinárodní konference, 17.–18. 6. 2014. Praha: ÚJOP UK, 140–148.